PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C07H 21/04

A1

(11) International Publication Number: WO 97/10365

(43) International Publication Date: 20 March 1997 (20.03.97)

(21) International Application Number: PCT/US96/14839

(22) International Filing Date: 13 September 1996 (13.09.96)

(30) Priority Data:
08/529,115 15 September 1995 (15.09.95) US

(71) Applicant *(for all designated States except US)*: AFFYMAX TECHNOLOGIES N.V. [NL/NL]; De Ruyderkade 62, Curaçao (AN).

(72) Inventors; and

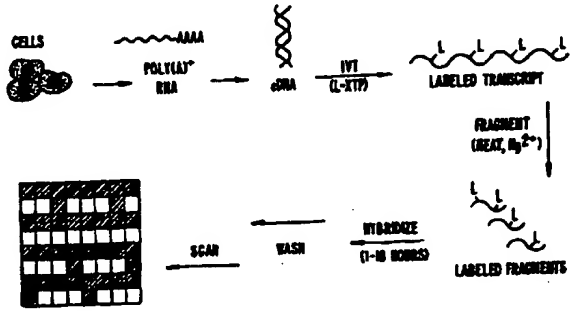
(75) Inventors/Applicants *(for US only)*: LOCKHART, David, J. [US/US]; 610 Mountain View Avenue, Mountain View, CA 94041 (US). BROWN, Eugene, L. [US/US]; 1388 Walnut Street, Newton Highlands, MA 02161 (US). WONG, Gordon [US/US]; 239 Clark Road, Brookline, MA 02146 (US). CHEE, Mark [AU/US]; 3199 Waverly Street, Palo Alto, CA 94306 (US). GINGERAS, Thomas, R. [US/US]; 528 Juniper Hill Drive, Encinitas, CA 92021 (US). MITTMANN, Michael, P. [US/US]; 2377 St. Francis Drive, Palo Alto, CA 94303 (US). LIPSHUTZ, Robert, J. [US/US]; 970 Palo Alto Avenue, Palo Alto, CA 94301 (US). FODOR, Stephen, P., A. [US/US]; 3863 Nathan Way, Palo Alto, CA 94303 (US). WANG, Chunwei

(74) Agents: HUNTER, Tom et al.; Townsend and Townsend and Crew L.L.P., 8th floor, Two Embarcadero Center, San Francisco, CA 94111-3834 (US).

(81) Designated States: AU, CA, JP, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: EXPRESSION MONITORING BY HYBRIDIZATION TO HIGH DENSITY OLIGONUCLEOTIDE ARRAYS



(57) Abstract

This invention provides methods of monitoring the expression levels of a multiplicity of genes. The methods involve hybridizing a nucleic acid sample to a high density array of oligonucleotide probes where the high density array contains oligonucleotide probes complementary to subsequences of target nucleic acids in the nucleic acid sample. In one embodiment, the method involves providing a pool of target nucleic acids comprising RNA transcripts of one or more target genes, or nucleic acids derived from the RNA transcripts, hybridizing said pool of nucleic acids to an array of oligonucleotide probes immobilized on surface, where the array comprising more than 100 different oligonucleotides and each different oligonucleotide is localized in a predetermined region of the surface, the density of the different oligonucleotides is greater than about 60 different oligonucleotides per 1 cm^2, and the oligonucleotide probes are complementary to the RNA transcripts or nucleic acids derived from the RNA transcripts; and quantifying the hybridized nucleic acids in the array.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia
AT	Austria
AU	Australia
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CS	Czechoslovakia
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgystan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakhstan
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam

EXPRESSION MONITORING BY HYBRIDIZATION TO HIGH DENSITY OLIGONUCLEOTIDE ARRAYS

CROSS REFERENCE TO RELATED APPLICATIONS

This is a continuation-in-part of U.S.S.N. 08/529,115 filed on September 15, 1995 which is herein incorporated by reference for all purposes.

BACKGROUND OF THE INVENTION

Many disease states are characterized by differences in the expression levels of various genes either through changes in the copy number of the genetic DNA or through changes in levels of transcription (*e.g.* through control of initiation, provision of RNA precursors, RNA processing, *etc.*) of particular genes. For example, losses and gains of genetic material play an important role in malignant transformation and progression. These gains and losses are thought to be "driven" by at least two kinds of genes. Oncogenes are positive regulators of tumorgenesis, while tumor suppressor genes are negative regulators of tumorgenesis (Marshall, *Cell*, 64: 313-326 (1991); Weinberg, Science, 254: 1138-1146 (1991)). Therefore, one mechanism of activating unregulated growth is to increase the number of genes coding for oncogene proteins or to increase the level of expression of these oncogenes (*e.g.* in response to cellular or environmental changes), and another is to lose genetic material or to decrease the level of expression of genes that code for tumor suppressors. This model is supported by the losses and gains of genetic material associated with glioma progression (Mikkelson *et al.* J. Cellular Biochm. 46: 3-8 (1991)). Thus, changes in the expression (transcription) levels of

particular genes (*e.g.* oncogenes or tumor suppressors), serve as signposts for the presence and progression of various cancers.

Similarly, control of the cell cycle and cell development, as well as diseases, are characterized by the variations in the transcription levels of particular genes. Thus, for example, a viral infection is often characterized by the elevated expression of genes of the particular virus. For example, outbreaks of *Herpes simplex*, Epstein-Barr virus infections (*e.g.* infectious mononucleosis), cytomegalovirus, Varicella-zoster virus infections, parvovirus infections, human papillomavirus infections, *etc.* are all characterized by elevated expression of various genes present in the respective virus. Detection of elevated expression levels of characteristic viral genes provides an effective diagnostic of the disease state. In particular, viruses such as herpes simplex, enter quiescent states for periods of time only to erupt in brief periods of rapid replication. Detection of expression levels of characteristic viral genes allows detection of such active proliferative (and presumably infective) states.

Oligonucleotide probes have long been used to detect complementary nucleic acid sequences in a nucleic acid of interest (the "target" nucleic acid) and have been used to detect expression of particular genes (*e.g.*, a Northern Blot). In some assay formats, the oligonucleotide probe is tethered, *i.e.*, by covalent attachment, to a solid support, and arrays of oligonucleotide probes immobilized on solid supports have been used to detect specific nucleic acid sequences in a target nucleic acid. *See, e.g.*, PCT patent publication Nos. WO 89/10977 and 89/11548. Others have proposed the use of large numbers of oligonucleotide probes to provide the complete nucleic acid sequence of a target nucleic acid but failed to provide an enabling method for using arrays of immobilized probes for this purpose. *See* U.S. Patent Nos. 5,202,231 and 5,002,867 and PCT patent publication No. WO 93/17126.

The use of "traditional" hybridization protocols for monitoring or quantifying gene expression is problematic. For example two or more gene products of approximately the same molecular weight will prove difficult or impossible to distinguish in a Northern blot because they are not readily separated by electrophoretic methods.

Similarly, as hybridization efficiency and cross-reactivity varies with the particular subsequence (region) of a gene being probed it is difficult to obtain an accurate and reliable measure of gene expression with one, or even a few, probes to the target gene.

The development of VLSIPS™ technology provided methods for synthesizing arrays of many different oligonucleotide probes that occupy a very small surface area. *See* U.S. Patent No. 5,143,854 and PCT patent publication No. WO 90/15070. U.S. Patent application Serial No. 082,937, filed June 25, 1993, describes methods for making arrays of oligonucleotide probes that can be used to provide the complete sequence of a target nucleic acid and to detect the presence of a nucleic acid containing a specific nucleotide sequence.

Prior to the present invention, however, it was unknown that high density oligonucleotide arrays could be used to reliably monitor message levels of a multiplicity of preselected genes in the presence of a large abundance of other (non-target) nucleic acids (*e.g.*, in a cDNA library, DNA reverse transcribed from an mRNA, mRNA used directly or amplified, or polymerized from a DNA template). In addition, the prior art provided no rapid and effective method for identifying a set of oligonucleotide probes that maximize specific hybridization efficacy while minimizing cross-reactivity nor of using hybridization patterns (in particular hybridization patterns of a multiplicity of oligonucleotide probes in which multiple oligonucleotide probes are directed to each target nucleic acid) for quantification of target nucleic acid concentrations.

Summary of the Invention

The present invention is premised, in part, on the discovery that microfabricated arrays of large numbers of different oligonucleotide probes (DNA chips) may effectively be used to not only detect the presence or absence of target nucleic acid sequences, but to quantify the relative abundance of the target sequences in a complex nucleic acid pool. In addition, it was also a surprising discovery that relatively short oligonucleotide probes (*e.g.*, 20 mer) are sufficiently specific to allow quantitation of gene expression in complex mixtures of nucleic acids particularly when provided as in high density oligonucleotide probe arrays.

Prior to this invention it was unknown that hybridization to high density probe arrays would permit small variations in expression levels of a particular gene to be identified and quantified in a complex population of nucleic acids that out number the target nucleic acids by 1,000 fold to 1,000,000 fold or more. It was also unknown that the transcription levels of specific genes can be quantitated in a complex nucleic acid mixture with only a few (*e.g.*, less than 20 or even less than 10) relatively short oligonucleotide probes.

Thus, this invention provides for a method of simultaneously monitoring the expression (*e.g.* detecting and or quantifying the expression) of a multiplicity of genes. The levels of transcription for virtually any number of genes may be determined simultaneously. Typically, at least about 10 genes, preferably at least about 100, more preferably at least about 1000 and most preferably at least about 10,000 different genes are assayed at one time.

The method involves providing a pool of target nucleic acids comprising mRNA transcripts of one or more of said genes, or nucleic acids derived from the mRNA transcripts; hybridizing the pool of nucleic acids to an array of oligonucleotide probes immobilized on a surface, where the array comprises more than 100 different oligonucleotides, each different oligonucleotide is localized in a predetermined region of said surface, each different oligonucleotide is attached to the surface through a single covalent bond, the density of the different oligonucleotides is greater than about 60 different oligonucleotides (where different oligonucleotides refers to oligonucleotides having different sequences) per 1 cm^2, and the oligonucleotide probes are complementary to the mRNA transcripts or nucleic acids derived from the mRNA transcripts; and quantifying the hybridized nucleic acids in the array. The method can additionally include a step of quantifying the hybridization of the target nucleic acids to the array. The quantification preferably provides a measure of the levels of transcription of the genes. In a preferred embodiment, the pool of target nucleic acids is one in which the concentration of the target nucleic acids (mRNA transcripts or nucleic acids derived from the mRNA transcripts) is proportional to the expression levels of genes encoding those target nucleic acids.

In a preferred embodiment, the array of oligonucleotide probes is a high density array comprising greater than about 100, preferably greater than about 1,000 more preferably greater than about 16,000 and most preferably greater than about 65,000 or 250,000 or even 1,000,000 different oligonucleotide probes. Such high density arrays comprise a probe density of generally greater than about 60, more generally greater than about 100, most generally greater than about 600, often greater than about 1000, more often greater than about 5,000, most often greater than about 10,000, preferably greater than about 40,000 more preferably greater than about 100,000, and most preferably greater than about 400,000 different oligonucleotide probes per cm^2 (where different oligonucleotides refers to oligonucleotides having different sequences). The oligonucleotide probes range from about 5 to about 50 nucleotides, preferably from about 5 to about 45 nucleotides, still more preferably from about 10 to about 40 nucleotides and most preferably from about 15 to about 40 nucleotides in length. Particularly preferred arrays contain probes ranging from about 20 to about 25 oligonucleotides in length. The array may comprise more than 10, preferably more than 50, more preferably more than 100, and most preferably more than 1000 oligonucleotide probes specific for each target gene. In a preferred embodiment, the array comprises at least 10 different oligonucleotide probes for each gene. In another preferred embodiment, the array 20 or fewer oligonucleotides complementary each gene. Although a planar array surface is preferred, the array may be fabricated on a surface of virtually any shape or even a multiplicity of surfaces.

The array may further comprise mismatch control probes. Where such mismatch controls are present, the quantifying step may comprise calculating the difference in hybridization signal intensity between each of the oligonucleotide probes and its corresponding mismatch control probe. The quantifying may further comprise calculating the average difference in hybridization signal intensity between each of the oligonucleotide probes and its corresponding mismatch control probe for each gene.

The probes present in the high density array can be oligonucleotide probes selected according to selection and optimization methods described below. Alternatively, non-optimal probes may be included in the array, but the probes used for

quantification (analysis) can be selected according to the optimization methods described below.

Oligonucleotide arrays for the practice of this invention are preferably chemically synthesized by parallel immobilized polymer synthesis methods, more preferably by light directed polymer synthesis methods. Chemically synthesized arrays are advantageous in that probe preparation does not require cloning, a nucleic acid amplification step, or enzymatic synthesis. Indeed, the preparation of the probes does not require handling of any biological materials.

The array includes test probes which are oligonucleotide probes each of which has a sequence that is complementary to a subsequence of one of the genes (or the mRNA or the corresponding antisense cRNA) whose expression is to be detected. In addition, the array can contain normalization controls, mismatch controls and expression level controls as described herein.

In a particularly preferred embodiment, the variation between different copies (within and/or between batches) of each array is less than 20%, more preferably less than about 10%, and most preferably less than about 5% where the variation is measured as the coefficient of variation in hybridization intensity averaged over at least 5 oligonucleotide probes for each gene whose expression the array is to detect.

The pool of nucleic acids may be labeled before, during, or after hybridization, although in a preferred embodiment, the nucleic acids are labeled before hybridization. Fluorescence labels are particularly preferred, more preferably labeling with a single fluorophore, and, where fluorescence labeling is used, quantification of the hybridized nucleic acids is by quantification of fluorescence from the hybridized fluorescently labeled nucleic acid. Such quantification is facilitated by the use of a fluorescence microscope which can be equipped with an automated stage to permit automatic scanning of the array, and which can be equipped with a data acquisition system for the automated measurement recording and subsequent processing of the fluorescence intensity information.

In a preferred embodiment, hybridization is at low stringency (*e.g.* about 20°C to about 50°C, more preferably about 30°C to about 40°C, and most preferably about 37°C and 6X SSPE-T or lower) with at least one wash at higher stringency.

Hybridization may include subsequent washes at progressively increasing stringency until a desired level of hybridization specificity is reached.

Quantification of the hybridization signal can be by any means known to one of skill in the art. However, in a particularly preferred embodiment, quantification is achieved by use of a confocal fluorescence microscope. Data is preferably evaluated by calculating the difference in hybridization signal intensity between each oligonucleotide probe and its corresponding mismatch control probe. It is particularly preferred that this difference be calculated and evaluated for each gene. Particularly preferred analytical methods are provided herein.

The pool of target nucleic acids can be the total polyA$^+$ mRNA isolated from a biological sample, or cDNA made by reverse transcription of the RNA or second strand cDNA or RNA transcribed from the double stranded cDNA intermediate. Alternatively, the pool of target nucleic acids can be treated to reduce the complexity of the sample and thereby reduce the background signal obtained in hybridization. In one approach, a pool of mRNAs, derived from a biological sample, is hybridized with a pool of oligonucleotides comprising the oligonucleotide probes present in the high density array. The pool of hybridized nucleic acids is then treated with RNase A which digests the single stranded regions. The remaining double stranded hybridization complexes are then denatured and the oligonucleotide probes are removed, leaving a pool of mRNAs enhanced for those mRNAs complementary to the oligonucleotide probes in the high density array.

In another approach to background reduction, a pool of mRNAs derived from a biological sample is hybridized with paired target specific oligonucleotides where the paired target specific oligonucleotides are complementary to regions flanking subsequences of the mRNAs complementary to the oligonucleotide probes in the high density array. The pool of hybridized nucleic acids is treated with RNase H which digests the hybridized (double stranded) nucleic acid sequences. The remaining single stranded nucleic acid sequences which have a length about equivalent to the region flanked by the paired target specific oligonucleotides are then isolated (*e.g.* by electrophoresis) and used as the pool of nucleic acids for monitoring gene expression.

Finally, a third approach to background reduction involves eliminating or reducing the representation in the pool of particular preselected target mRNA messages (*e.g.*, messages that are characteristically overexpressed in the sample). This method involves hybridizing an oligonucleotide probe that is complementary to the preselected target mRNA message to the pool of polyA$^+$ mRNAs derived from a biological sample. The oligonucleotide probe hybridizes with the particular preselected polyA$^+$ mRNA (message) to which it is complementary. The pool of hybridized nucleic acids is treated with RNase H which digests the double stranded (hybridized) region thereby separating the message from its polyA$^+$ tail. Isolating or amplifying (*e.g.*, using an oligo dT column) the polyA$^+$ mRNA in the pool then provides a pool having a reduced or no representation of the preselected target mRNA message.

It will be appreciated that the methods of this invention can be used to monitor (detect and/or quantify) the expression of any desired gene of known sequence or subsequence. Moreover, these methods permit monitoring expression of a large number of genes simultaneously and effect significant advantages in reduced labor, cost and time. The simultaneous monitoring of the expression levels of a multiplicity of genes permits effective comparison of relative expression levels and identification of biological conditions characterized by alterations of relative expression levels of various genes. Genes of particular interest for expression monitoring include genes involved in the pathways associated with various pathological conditions (*e.g.*, cancer) and whose expression is thus indicative of the pathological condition. Such genes include, but are not limited to the HER2 (c-erbB-2/neu) proto-oncogene in the case of breast cancer, receptor tyrosine kinases (RTKs) associated with the etiology of a number of tumors including carcinomas of the breast, liver, bladder, pancreas, as well as glioblastomas, sarcomas and squamous carcinomas, and tumor suppressor genes such as the P53 gene and other "marker" genes such as RAS, MSH2, MLH1 and BRCA1. Other genes of particular interest for expression monitoring are genes involved in the immune response (*e.g.*, interleukin genes), as well as genes involved in cell adhesion (*e.g.*, the integrins or selectins) and signal transduction (*e.g.*, tyrosine kinases), *etc.*

In another embodiment, this invention provides a method of identifying genes that are effected by one or more drugs, or conversely, screening a number of

drugs to identify those that have an effect on particular gene(s). This involves providing a pool of target nucleic acids from one or more cells contacted with the drug or drugs and hybridizing that pool to any of the high density oligonucleotide arrays described herein. The expression levels of the genes targeted by the probes in the array are determined and compared to expression levels of genes from "control" cells not exposed to the drug or drugs. The genes that are overexpressed or underexpressed in response to the drug or drugs are identified or conversely the drug or drugs that alter expression of one or more genes are identified.

In still yet another embodiment, this invention provide for a composition comprising any of the high density oligonucleotide arrays disclosed herein where the oligonucleotide probes are specifically hybridized to one or more fluorescently labeled nucleic acids (which are the transcription products of genes or derived from those transcription products) thereby forming a fluorescent array in which the fluorescence of the array is indicative of the transcription levels of the multiplicity of genes. One of skill will appreciate that such a hybridized array may be used as a reference, control, or standard (*e.g.*, provided in a kit) or may itself be a diagnostic array indicating the expression levels of a multiplicity of genes in a sample.

This invention also provides kits for simultaneously monitoring expression levels of a multiplicity of genes. The kits include an array of immobilized oligonucleotide probes complementary to subsequences of the multiplicity of target genes, as described herein. The kit may also include instructions describing the use of the array for detection and/or quantification of expression levels of the multiplicity of genes. The kit may additionally include one or more of the following: buffers, hybridization mix, wash and read solutions, labels, labeling reagents (enzymes *etc.*), "control" nucleic acids, software for probe selection, array reading or data analysis and any of the other materials or reagents described herein for the practice of the claimed methods.

In another embodiment, this invention provides for a method of selecting a set of oligonucleotide probes, that specifically bind to a target nucleic acid (*e.g.*, a gene or genes whose expression is to be monitored or nucleic acids derived from the gene or its transcribed mRNA). The method involves providing a high density array of

oligonucleotide probes where the array comprises a multiplicity of probes wherein each probe is complementary to a subsequence of the target nucleic acid. The target nucleic acid is then hybridized to the array of oligonucleotide probes to identify and select those probes where the difference in hybridization signal intensity between each probe and its mismatch control is detectable (preferably greater than about 10% of the background signal intensity, more preferably greater than about 20% of the background signal intensity and most preferably greater than about 50% of the background signal intensity). The method can further comprise hybridizing the array to a second pool of nucleic acids comprising nucleic acids other than the target nucleic acids; and identifying and selecting probes having the lowest hybridization signal and where both the probe and its mismatch control have a hybridization intensity equal to or less than about 5 times the background signal intensity, preferably equal to or less than about 2 times the background signal intensity, more preferably equal to or less than about 1 times the background signal intensity, and most preferably equal or less than about half the background signal intensity.

In a preferred embodiment, the multiplicity of probes can include every different probe of length n that is complementary to a subsequence of the target nucleic acid. The probes can range from about 10 to about 50 nucleotides in length. The array is preferably a high density array as described above. Similarly, the hybridization methods, conditions, times, fluid volumes, detection methods are as herein .

In another embodiment, the invention provides a computer-implemented method of monitoring expression of genes comprising the steps of: receiving input of hybridization intensities for a plurality of nucleic acid probes including pairs of perfect match probes and mismatch probes, the hybridization intensities indicating hybridization affinity between the plurality of nucleic acid probes and nucleic acids corresponding to a gene, and each pair including a perfect match probe that is perfectly complementary to a portion of the nucleic acids and a mismatch probe that differs from the perfect match probe by at least one nucleotide; comparing the hybridization intensities of the perfect match and mismatch probes of each pair; and indicating expression of the gene according to results of the comparing step. Preferably, the differences between the

hybridization intensities of the perfect match and mismatch probes of each pair are calculated.

Additionally, the invention provides a computer-implemented method for monitoring expression of genes comprising the steps of: receiving input of a nucleic acid sequence constituting a gene; generating a set of probes that are perfectly complementary to the gene; and identifying a subset of probes, including less than all of the probes in the set, for monitoring the expression of the gene. Each probe of the set may be analyzed by criteria that specify characteristics indicative of low hybridization or high cross hybridization. The criteria may include if occurrences of a specific nucleotide in a probe crosses a threshold value, if the number of a specific nucleotide that repeats sequentially in a probe crosses a threshold value, if the length of a palindrome in a probe crosses a threshold value, and the like.

Definitions.

The phrase "massively parallel screening" refers to the simultaneous screening of at least about 100, preferably about 1000, more preferably about 10,000 and most preferably about 1,000,000 different nucleic acid hybridizations.

The terms "nucleic acid" or "nucleic acid molecule" refer to a deoxyribonucleotide or ribonucleotide polymer in either single-or double-stranded form, and unless otherwise limited, would encompass known analogs of natural nucleotides that can function in a similar manner as naturally occurring nucleotides.

An oligonucleotide is a single-stranded nucleic acid ranging in length from 2 to about 500 bases.

As used herein a "probe" is defined as an oligonucleotide capable of binding to a target nucleic acid of complementary sequence through one or more types of chemical bonds, usually through complementary base pairing, usually through hydrogen bond formation. As used herein, an oligonucleotide probe may include natural (*i.e.* A, G, C, or T) or modified bases (7-deazaguanosine, inosine, *etc.*). In addition, the bases in oligonucleotide probe may be joined by a linkage other than a phosphodiester bond, so long as it does not interfere with hybridization. Thus, oligonucleotide probes may be

peptide nucleic acids in which the constituent bases are joined by peptide bonds rather than phosphodiester linkages.

The term "target nucleic acid" refers to a nucleic acid (often derived from a biological sample), to which the oligonucleotide probe is designed to specifically hybridize. It is either the presence or absence of the target nucleic acid that is to be detected, or the amount of the target nucleic acid that is to be quantified. The target nucleic acid has a sequence that is complementary to the nucleic acid sequence of the corresponding probe directed to the target. The term target nucleic acid may refer to the specific subsequence of a larger nucleic acid to which the probe is directed or to the overall sequence (*e.g.*, gene or mRNA) whose expression level it is desired to detect. The difference in usage will be apparent from context.

"Subsequence" refers to a sequence of nucleic acids that comprise a part of a longer sequence of nucleic acids.

The term "complexity"is used here according to standard meaning of this term as established by Britten *et al. Methods of Enzymol.* 29:363 (1974). See, also *Cantor and Schimmel Biophysical Chemistry: Part III* at 1228-1230 for further explanation of nucleic acid complexity.

"Bind(s) substantially" refers to complementary hybridization between a probe nucleic acid and a target nucleic acid and embraces minor mismatches that can be accommodated by reducing the stringency of the hybridization media to achieve the desired detection of the target polynucleotide sequence.

The phrase "hybridizing specifically to", refers to the binding, duplexing, or hybridizing of a molecule only to a particular nucleotide sequence under stringent conditions when that sequence is present in a complex mixture (*e.g.*, total cellular) DNA or RNA. The term "stringent conditions" refers to conditions under which a probe will hybridize to its target subsequence, but to no other sequences. Stringent conditions are sequence-dependent and will be different in different circumstances. Longer sequences hybridize specifically at higher temperatures. Generally, stringent conditions are selected to be about 5°C lower than the thermal melting point (Tm) for the specific sequence at a defined ionic strength and pH. The Tm is the temperature (under defined ionic strength, pH, and nucleic acid concentration) at which 50% of the probes

complementary to the target sequence hybridize to the target sequence at equilibrium. (As the target sequences are generally present in excess, at Tm, 50% of the probes are occupied at equilibrium). Typically, stringent conditions will be those in which the salt concentration is at least about 0.01 to 1.0 M Na ion concentration (or other salts) at pH 7.0 to 8.3 and the temperature is at least about 30°C for short probes (*e.g.*, 10 to 50 nucleotides). Stringent conditions may also be achieved with the addition of destabilizing agents such as formamide.

The term "perfect match probe" refers to a probe that has a sequence that is perfectly complementary to a particular target sequence. The test probe is typically perfectly complementary to a portion (subsequence) of the target sequence. The perfect match (PM) probe can be a "test probe", a "normalization control" probe, an expression level control probe and the like. A perfect match control or perfect match probe is, however, distinguished from a "mismatch control" or "mismatch probe."

The term "mismatch control" or "mismatch probe" refer to probes whose sequence is deliberately selected not to be perfectly complementary to a particular target sequence. For each mismatch (MM) control in a high-density array there typically exists a corresponding perfect match (PM) probe that is perfectly complementary to the same particular target sequence. The mismatch may comprise one or more bases. While the mismatch(s) may be locates anywhere in the mismatch probe, terminal mismatches are less desirable as a terminal mismatch is less likely to prevent hybridization of the target sequence. In a particularly preferred embodiment, the mismatch is located at or near the center of the probe such that the mismatch is most likely to destabilize the duplex with the target sequence under the test hybridization conditions.

The terms "background" or "background signal intensity" refer to hybridization signals resulting from non-specific binding, or other interactions, between the labeled target nucleic acids and components of the oligonucleotide array (*e.g.*, the oligonucleotide probes, control probes, the array substrate, *etc.*). Background signals may also be produced by intrinsic fluorescence of the array components themselves. A single background signal can be calculated for the entire array, or a different background signal may be calculated for each target nucleic acid. In a preferred embodiment, background is calculated as the average hybridization signal intensity for the lowest 5%

to 10% of the probes in the array, or, where a different background signal is calculated for each target gene, for the lowest 5% to 10% of the probes for each gene. Of course, one of skill in the art will appreciate that where the probes to a particular gene hybridize well and thus appear to be specifically binding to a target sequence, they should not be used in a background signal calculation. Alternatively, background may be calculated as the average hybridization signal intensity produced by hybridization to probes that are not complementary to any sequence found in the sample (*e.g.* probes directed to nucleic acids of the opposite sense or to genes not found in the sample such as bacterial genes where the sample is mammalian nucleic acids). Background can also be calculated as the average signal intensity produced by regions of the array that lack any probes at all.

The term "quantifying" when used in the context of quantifying transcription levels of a gene can refer to absolute or to relative quantification. Absolute quantification may be accomplished by inclusion of known concentration(s) of one or more target nucleic acids (*e.g.* control nucleic acids such as Bio B or with known amounts the target nucleic acids themselves) and referencing the hybridization intensity of unknowns with the known target nucleic acids (*e.g.* through generation of a standard curve). Alternatively, relative quantification can be accomplished by comparison of hybridization signals between two or more genes, or between two or more treatments to quantify the changes in hybridization intensity and, by implication, transcription level.

The"percentage of sequence identity" or "sequence identity" is determined by comparing two optimally aligned sequences or subsequences over a comparison window or span, wherein the portion of the polynucleotide sequence in the comparison window may optionally comprise additions or deletions (*i.e.*, gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical subunit (*e.g.* nucleic acid base or amino acid residue) occurs in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Percentage sequence identity when calculated using the programs GAP or BESTFIT (see below) is calculated using default gap weights.

Methods of alignment of sequences for comparison are well known in the art. Optimal alignment of sequences for comparison may be conducted by the local homology algorithm of Smith and Waterman, *Adv. Appl. Math.* 2: 482 (1981), by the homology alignment algorithm of Needleman and Wunsch *J. Mol. Biol.* 48: 443 (1970), by the search for similarity method of Pearson and Lipman, *Proc. Natl. Acad. Sci. USA* 85: 2444 (1988), by computerized implementations of these algorithms (including, but not limited to CLUSTAL in the PC/Gene program by Intelligenetics, Moutain View, California, GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group (GCG), 575 Science Dr., Madison, Wisconsin, USA), or by inspection. In particular, methods for aligning sequences using the CLUSTAL program are well described by Higgins and Sharp in *Gene*, 73: 237-244 (1988) and in *CABIOS* 5: 151-153 (1989)).

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 shows a schematic of expression monitoring using oligonucleotide arrays. Extracted poly $(A)^+$ RNA is converted to cDNA, which is then transcribed in the presence of labeled ribonucleotide triphosphates. L is either biotin or a dye such as fluorescein. RNA is fragmented with heat in the presence of magnesium ions. Hybridizations are carried out in a flow cell that contains the two-dimensional DNA probe arrays. Following a brief washing step to remove unhybridized RNA, the arrays are scanned using a scanning confocal microscope. Alternatives in which cellular mRNA is directly labeled without a cDNA intermediate are described in the Examples. Image analysis software converts the scanned array images into text files in which the observed intensities at specific physical locations are associated with particular probe sequences.

Fig. 2A shows a fluorescent image of a high density array containing over 16,000 different oligonucleotide probes. The image was obtained following hybridization (15 hours at 40°C) of biotin-labeled randomly fragmented sense RNA transcribed from the murine B cell (Tl0) cDNA library, and spiked at the level of 1:3,000 (50 pM equivalent to about 100 copies per cell) with 13 specific RNA targets. The brightness at any location is indicative of the amount of labeled RNA hybridized to the particular oligonucleotide probe. Fig. 2B shows a small portion of the array (the boxed region of Fig. 2A) containing probes

for IL-2 and IL-3 RNAS, For comparison, Fig. 2C shows shown the same region of the array following hybridization with an unspiked T10 RNA samples (T10 cells do not express IL-2 and IL-3). The variation in the signal intensity was highly reproducible and reflected the sequence dependence of the hybridization efficiencies. The central cross and the four corners of the array contain a control sequence that is complementary to a biotin-labeled oligonucleotide that was added to the hybridization solution at a constant concentration (50 pM). The sharpness of the images near the boundaries of the features was limited by the resolution of the reading device (11.25 μm) and not by the spatial resolution of the array synthesis. The pixels in the border regions of each synthesis feature were systematically ignored in the quantitative analysis of the images.

Fig. 3 provides a log/log plot of the hybridization intensity (average of the PM-MM intensity differences for each gene) versus concentration for 11 different RNA targets. The hybridization signals were quantitatively related to target concentration. The experiments were performed as described in the Examples herein and in Fig. 2. The ten 10 cytokine RNAs (plus *bioB*) were spiked into labeled T10 RNA at levels ranging from 1:300,000 to 1:3,000. The signals continued to increase with increased concentration up to frequencies of 1:300, but the response became sublinear at the high levels due to saturation of the probe sites, The linear range can be extended to higher concentrations by using shorter hybridization times. RNAs from genes expressed in T10 cells (IL-10, β-actin and GAPDH) were also detected at levels consistent with results obtained by probing cDNA libraries.

Fig. 4 shows cytokine mRNA levels in the murine 2D6 T helper cell line at different times following stimulation with PMA and a calcium ionophore. Poly (A)$^+$ RNA was extracted at 0, 2, 6, and 24 hours following stimulation and converted to double stranded cDNA containing an RNA polymerase promoter. The cDNA pool was then transcribed in the presence of biotin labeled ribonucleotide triphosphates, fragmented, and hybridized to the oligonucleotide probe arrays for 2 and 22 hours. The fluorescence intensities were converted to RNA frequencies by comparison with the signals obtained for a bacterial RNA (biotin synthetase) spiked into the samples at known amounts prior to hybridization. A signal of 50,000 corresponds to a frequency of approximately 1:100,000 to a frequency of 1:5,000, and a signal of 100 to a frequency of 1:50,000. RNAs for IL-2,

IL-4, IL-6, and IL-12p40 were not detected above the level of approximately 1:200,000 in these experiments. The error bars reflect the estimated uncertainty (25 percent) in the level for a given RNA relative to the level for the same RNA at a different time point. The relative uncertainty estimate was based on the results of repeated spiking experiments, and on repeated measurements of IL-10, β-actin and GAPDH RNAs in preparations from both T10 and 2D6 cells (unstimulated). The uncertainty in the absolute frequencies includes message-to-message differences in the hybridization efficiency as well as differences in the mRNA isolation, cDNA synthesis, and RNA synthesis and labeling steps. The uncertainty in the absolute frequencies is estimated to be a factor of three.

Fig. 5 shows a fluorescence image of an array containing over 63,000 different oligonucleotide probes for 118 genes. The image was obtained following overnight hybridization of a labeled murine B cell RNA sample. Each square synthesis region is 50 x 50 μm and contains 107 to 108 copies of a specific oligonucleotide. The array was scanned at a resolution of 7.5 μm in approximately 15 minutes. The bright rows indicate RNAs present at high levels. Lower level RNAs were unambiguously detected based on quantitative evaluation of the hybridization patterns. A total of 21 murine RNAs were detected at levels ranging from approximately 1:300,000 to 1:100. The cross in the center, the checkerboard in the corners, and the MUR-1 region at the top contain probes complementary to a labeled control oligonucleotide that was added to all samples.

Fig. 6 shows an example of a computer system used to execute the software of an embodiment of the present invention.

Fig. 7 shows a system block diagram of a typical computer system used to execute the software of an embodiment of the present invention.

Fig. 8 shows the high level flow of a process of monitoring the expression of a gene by comparing hybridization intensities of pairs of perfect match and mismatch probes.

Fig. 9 shows the flow of a process of determining if a gene is expressed utilizing a decision matrix.

Figs. 10A and 10B show the flow of a process of determining the expression of a gene by comparing baseline scan data and experimental scan data.

Fig. 11 shows the flow of a process of increasing the number of probes for monitoring the expression of genes after the number of probes has been reduced or pruned.

DETAILED DESCRIPTION

I. High Density Arrays For Monitoring Gene Expression

This invention provides methods of monitoring (detecting and/or quantifying) the expression levels of one or more genes. The methods involve hybridization of a nucleic acid target sample to a high density array of nucleic acid probes and then quantifying the amount of target nucleic acids hybridized to each probe in the array.

While nucleic acid hybridization has been used for some time to determine the expression levels of various genes (*e.g.*, Northern Blot), it was a surprising discovery of this invention that high density arrays are suitable for the quantification of the small variations in expression (transcription) levels of a gene in the presence of a large population of heterogenous nucleic acids. The signal may be present at a concentration of less than about 1 in 1,000, and is often present at a concentration less than 1 in 10,000 more preferably less than about 1 in 50,000 and most preferably less than about 1 in 100,000, 1 in 300,000, or even 1 in 1,000,000.

Prior to this invention, it was expected that hybridization of such a complex mixture to a high density array might overwhelm the available probes and make it impossible to detect the presence of low-level target nucleic acids. It was thus unclear that a low level signal could be isolated and detected in the presence of misleading signals due to cross-hybridization and non-specific binding both to substrate and probe. It was therefore a surprising discovery that, to the contrary, high density arrays are particularly well suited for monitoring expression of a multiplicity of genes and provide a level of sensitivity and discrimination hitherto unexpected.

It was also a surprising discovery of this invention that when used in a high-density array, even relatively short oligonucleotides can be used to accurately detect and quantify expression (transcription) levels of genes. Thus oligonucleotide arrays having oligonucleotides as short as 10 nucleotides, more preferably 15 oligonucleotides and most preferably 20 or 25 oligonucleotides are used to specifically detect and quantify

gene expression levels. Of course arrays containing longer oligonucleotides, as described herein, are also suitable.

A) Advantages of Oligonucleotide Arrays

In one preferred embodiment, the high density arrays used in the methods of this invention comprise chemically synthesized oligonucleotides. The use of chemically synthesized oligonucleotide arrays, as opposed to, for example, blotted arrays of genomic clones, restriction fragments, oligonucleotides, and the like, offers numerous advantages. These advantages generally fall into four categories:

1) Efficiency of production;
2) Reduced intra- and inter-array variability;
3) Increased information content; and
4) Higher signal to noise ratio (improved sensitivity).

1) Efficiency of production.

In a preferred embodiment, the arrays are synthesized using methods of spatially addressed parallel synthesis (*see, e.g.*, Section V, below). The oligonucleotides are synthesized chemically in a highly parallel fashion covalently attached to the array surface. This allows extremely efficient array production. For example, arrays containing tens (or even hundreds) of thousands of specifically selected 20 mer oligonucleotides are synthesized in fewer than 80 synthesis cycles. The arrays are designed and synthesized based on sequence information alone. Thus, unlike blotting methods, the array preparation requires no handling of biological materials. There is no need for cloning steps, nucleic acid amplifications, cataloging of clones or amplification products, and the like. The preferred chemical synthesis of expression monitoring arrays in this invention is thus more efficient blotting methods and permits the production of highly reproducible high-density arrays with relatively little labor and expense.

2) Reduced intra- and inter-array variability.

The use of chemically synthesized high-density oligonucleotide arrays in the methods of this invention improves intra- and inter-array variability. The

oligonucleotide arrays preferred for this invention are made in large batches (presently 49 arrays per wafer with multiple wafers synthesized in parallel) in a highly controlled reproducible manner. This makes them suitable as general diagnostic and research tools permitting direct comparisons of assays performed anywhere in the world.

Because of the precise control obtainable during the chemical synthesis the arrays of this invention show less than about 25%, preferably less than about 20%, more preferably less than about 15%, still more preferably less than about 10%, even more preferably less than about 5%. and most preferably less than about 2% variation between high density arrays (within or between production batches) having the same probe composition. Array variation is assayed as the variation in hybridization intensity (against a labeled control target nucleic acid mixture) in one or more oligonucleotide probes between two or more arrays. More preferably, array variation is assayed as the variation in hybridization intensity (against a labeled control target nucleic acid mixture) measured for one or more target genes between two or more arrays.

In addition to reducing inter- and intra-array variability, chemically synthesized arrays also reduce variations in relative probe frequency inherent in spotting methods, particularly spotting methods that use cell-derived nucleic acids (*e.g.*, cDNAs). Many genes are expressed at the level of thousands of copies per cell, while others are expressed at only a single copy per cell. A cDNA library will reflect this very large bias as will a cDNA library made from theis material. While normalization (adjustment of the amount of each different probe *e.g.*, by comparison to a reference cDNA) of the library will reduce the representation of over-expressed sequences, normalization has been shown to lessen the odds of selecting highly expressed cDNAs by only about a factor of 2 or 3. In contrast, chemical synthesis methods can insure that all oligonucleotide probes are represented in approximately equal concentrations. This decreases the inter-gene (intra-array) variability and permits direct comparison between characteristically overexpressed and underexpressed nucleic acids.

3) Increased information content.

As indicated above, it was a discovery of this invention that the use of high density oligonucleotide arrays for expression monitoring provides a number of

advantages not found with other methods. For example, the use of large numbers of different probes that specifically bind to the transcription product of a particular target gene provides a high degree of redundancy and internal control that permits optimization of probe sets for effective detection of particular target genes and minimizes the possibility of errors due to cross-reactivity with other nucleic acid species.

Apparently suitable probes often prove ineffective for expression monitoring by hybridization. For example, certain subsequences of a particular target gene may be found in other regions of the genome and probes directed to these subsequences will cross-hybridize with the other regions and not provide a signal that is a meaningful measure of the expression level of the target gene. Even probes that show little cross reactivity may be unsuitable because they generally show poor hybridization due to the formation of structures that prevent effective hybridization. Finally, in sets with large numbers of probes, it is difficult to identify hybridization conditions that are optimal for all the probes in a set. Because of the high degree of redundancy provided by the large number of probes for each target gene, it is possible to eliminate those probes that function poorly under a given set of hybridization conditions and still retain enough probes to a particular target gene to provide an extremely sensitive and reliable measure of the expression level (transcription level) of that gene.

In addition, the use of large numbers of different probes to each target gene makes it possible to monitor expression of families of closely-related nucleic acids. The probes may be selected to hybridize both with subsequences that are conserved across the family and with subsequences that differ in the different nucleic acids in the family. Thus, hybridization with such arrays permits simultaneous monitoring of the various members of a gene family even where the various genes are approximately the same size and have high levels of homology. Such measurements are difficult or impossible with traditional hybridization methods.

Because the high density arrays contain such a large number of probes it is possible to provide numerous controls including, for example, controls for variations or mutations in a particular gene, controls for overall hybridization conditions, controls for sample preparation conditions, controls for metabolic activity of the cell from which

the nucleic acids are derived and mismatch controls for non-specific binding or cross hybridization.

Moreover, as explained above, it was a surprising discovery of this invention that effective detection and quantitation of gene transcription in complex mammalian cell message populations can be determined with relatively short oligonucleotides and with relative few (*e.g.*, fewer than 40, preferably fewer than 30, more preferably fewer than 25, and most preferably fewer than 20, 15, or even 10) oligonucleotide probes per gene. In general, it was a discovery of this invention that there are a large number of probes which hybridize both strongly and specifically for each gene. This does not mean that a large number of probes is required for detection, but rather that there are many from which to choose and that choices can be based on other considerations such as sequence uniqueness (gene families), checking for splice variants, or genotyping hot spots (things not easily done with cDNA spotting methods).

Based on these discoveries, sets of four arrays are made that contain approximately 400,000 probes each. Sets of about 40 probes (20 probe pairs) are chosen that are complementary to each of about 40,000 genes for which there are ESTs in the public database. This set of ESTs covers roughly one-third to one-half of all human genes and these arrays will allow the levels of all of them to be monitored in a parallel set of overnight hybridizations.

4) Improved signal to noise ratio.

Blotted nucleic acids typically rely on ionic, electrostatic, and hydrophobic interactions to attach the blotted nucleic acids to the substrate. Bonds are formed at multiple points along the nucleic acid restricting degrees of freedom and interferign with the ability of the nucleic acid to hybridize to its complementary target. In contrast, the preferred arrays of this invention are chemically synthesized. The oligonucleotide probes are attached to the substrate by a single terminal covalent bond. The probes have more degrees of freedom and are capable of participating in complex interactions with their complementary targets. Consequently, the probe arrays of this invention show significantly higher hybridization efficiencies (10 times, 100 times, and even 1000 times more effecient) than blotted arrays. Less target oligonucleotide is used

to produce a given signal thereby dramatically improving the signal to noise ratio. Consequently the methods of this invention permit detection of only a few copies of a nucleic acid in extremely complex nucleic acid mixtures.

B) Preferred High Density Arrays

Preferred high density arrays of this invention comprise greater than about 100, preferably greater than about 1000, more preferably greater than about 16,000 and most preferably greater than about 65,000 or 250,000 or even greater than about 1,000,000 different oligonucleotide probes. The oligonucleotide probes range from about 5 to about 50 or about 5 to about 45 nucleotides, more preferably from about 10 to about 40 nucleotides and most preferably from about 15 to about 40 nucleotides in length. In particular preferred embodiments, the oligonucleotide probes are 20 or 25 nucleotides in length. It was a discovery of this invention that relatively short oligonucleotide probes sufficient to specifically hybridize to and distinguish target sequences. Thus in one preferred embodiment, the oligonucleotide probes are less than 50 nucleotides in length, generally less than 46 nucleotides, more generally less than 41 nucleotides, most generally less than 36 nucleotides, preferably less than 31 nucleotides, more preferably less than 26 nucleotides,and most preferably less than 21 nucleotides in length. The probes can also be less than 16 nucleotides or less than even 11 nucleotides in length.

The location and sequence of each different oligonucleotide probe sequence in the array is known. Moreover, the large number of different probes occupies a relatively small area providing a high density array having a probe density of generally greater than about 60, more generally greater than about 100, most generally greater than about 600, often greater than about 1000, more often greater than about 5,000, most often greater than about 10,000, preferably greater than about 40,000 more preferably greater than about 100,000, and most preferably greater than about 400,000 different oligonucleotide probes per cm^2. The small surface area of the array (often less than about 10 cm^2, preferably less than about 5 cm^2 more preferably less than about 2 cm^2, and most preferably less than about 1.6 cm^2) permits extremely uniform

hybridization conditions (temperature regulation, salt content, *etc.*) while the extremely large number of probes allows massively parallel processing of hybridizations.

Finally, because of the small area occupied by the high density arrays, hybridization may be carried out in extremely small fluid volumes (*e.g.*, 250 μl or less, more preferably 100 μl or less, and most preferably 10 μl or less). In small volumes, hybridization may proceed very rapidly. In addition, hybridization conditions are extremely uniform throughout the sample, and the hybridization format is amenable to automated processing.

II. Uses of Expression monitoring.

This invention demonstrates that hybridization with high density oligonucleotide probe arrays provides an effective means of monitoring expression of a multiplicity of genes. In addition this invention provides for methods of sample treatment and array designs and methods of probe selection that optimize signal detection at extremely low concentrations in complex nucleic acid mixtures.

The expression monitoring methods of this invention may be used in a wide variety of circumstances including detection of disease, identification of differential gene expression between two samples (*e.g.*, a pathological as compared to a healthy sample), screening for compositions that upregulate or downregulate the expression of particular genes, and so forth.

In one preferred embodiment, the methods of this invention are used to monitor the expression (transcription) levels of nucleic acids whose expression is altered in a disease state. For example, a cancer may be characterized by the overexpression of a particular marker such as the HER2 (c-erbB-2/neu) proto-oncogene in the case of breast cancer. Similarly, overexpression of receptor tyrosine kinases (RTKs) is associated with the etiology of a number of tumors including carcinomas of the breast, liver, bladder, pancreas, as well as glioblastomas, sarcomas and squamous carcinomas (*see* Carpenter, *Ann. Rev. Biochem.*, 56: 881-914 (1987)). Conversely, a cancer (*e.g.*, colerectal, lung and breast) may be characterized by the mutation of or underexpression of a tumor suppressor gene such as P53 (*see, e.g.*, Tominaga *et al. Critical Rev. in Oncogenesis*, 3: 257-282 (1992)).

In another preferred embodiment, the methods of this invention are used to monitor expression of various genes in response to defined stimuli, such as a drug. The methods are particularly advantageous because they permit simultaneous monitoring of the expression of thousands of genes. This is especially useful in drug research if the end point description is a complex one, not simply asking if one particular gene is overexpressed or underexpressed. Thus, where a disease state or the mode of action of a drug is not well characterized, the methods of this invention allow rapid determination of the particularly relevant genes.

As indicated above, the materials and methods of this invention are typically used to monitor the expression of a multiplicity of different genes simultaneously. Thus, in one embodiment, the invention provide for simultaneous monitoring of at least about 10, preferably at least about 100, more preferably at least about 1000, still more preferably at least about 10,000, and most preferably at least about 100,000 different genes.

The expression monitoring methods of this invention can also be used for gene discovery. Many genes that have been discovered to date have been classified into families based on commonality of the sequences. Because of the extremely large number of probes it is possible to place in the high density array, it is possible to include oligonucleotide probes representing known or parts of known members from every gene class. In utilizing such a "chip" (high density array) genes that are already known would give a positive signal at loci containing both variable and common regions. For unknown genes, only the common regions of the gene family would give a positive signal. The result would indicate the possibility of a newly discovered gene.

The expression monitoring methods of this invention also allow the development of "dynamic" gene databases. The Human Genome Project and commercial sequencing projects have generated large static databases which list thousands of sequences without regard to function or genetic interaction. Expression analysis using the methods of this invention produces "dynamic" databases that define a gene's function and its interactions with other genes. Without the ability to monitor the expression of large numbers of genes simultaneously ,however, the work of creating such a database is enormous. The tedious nature of using DNA sequence analysis for

determining an expression pattern involves preparing a cDNA library from the RNA isolated from the cells of interest and then sequencing the library. As the DNA is sequenced, the operator lists the sequences that are obtained and counts them. Thousands of sequences would have to be determined and then the frequency of those gene sequences would define the expression pattern of genes for the cells being studied.

By contrast, using an expression monitoring array to obtain the data according to the methods of this invention is relatively fast and easy. The process involves stimulating the cells to induce expression, obtaining the RNA from the cells and then either labeling the RNA directly or creating a cDNA copy of the RNA. If cDNA is to be hybridized to the chip, fluorescent molecules are incorporated during the DNA polymerization. Either the labeled RNA or the labeled cDNA is then hybridized to a high density array in one overnight experiment. The hybridization provides a quantitative assessment of the levels of every single one of the genes with no additional sequencing. In addition the methods of this invention are much more sensitive allowing a few copies of expressed genes per cell to be detected. This procedure is demonstrated in the examples provided herein.

III. Methods of monitoring gene expression.

Generally the methods of monitoring gene expression of this invention involve (1) providing a pool of target nucleic acids comprising RNA transcript(s) of one or more target gene(s), or nucleic acids derived from the RNA transcript(s); (2) hybridizing the nucleic acid sample to a high density array of probes (including control probes); and (3) detecting the hybridized nucleic acids and calculating a relative expression (transcription) level.

A) Providing a nucleic acid sample.

One of skill in the art will appreciate that in order to measure the transcription level (and thereby the expression level) of a gene or genes, it is desirable to provide a nucleic acid sample comprising mRNA transcript(s) of the gene or genes, or nucleic acids derived from the mRNA transcript(s). As used herein, a nucleic acid derived from an mRNA transcript refers to a nucleic acid for whose synthesis the mRNA

transcript or a subsequence thereof has ultimately served as a template. Thus, a cDNA reverse transcribed from an mRNA, an RNA transcribed from that cDNA, a DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, *etc.*, are all derived from the mRNA transcript and detection of such derived products is indicative of the presence and/or abundance of the original transcript in a sample. Thus, suitable samples include, but are not limited to, mRNA transcripts of the gene or genes, cDNA reverse transcribed from the mRNA, cRNA transcribed from the cDNA, DNA amplified from the genes, RNA transcribed from amplified DNA, and the like.

In a particularly preferred embodiment, where it is desired to quantify the transcription level (and thereby expression) of a one or more genes in a sample, the nucleic acid sample is one in which the concentration of the mRNA transcript(s) of the gene or genes, or the concentration of the nucleic acids derived from the mRNA transcript(s), is proportional to the transcription level (and therefore expression level) of that gene. Similarly, it is preferred that the hybridization signal intensity be proportional to the amount of hybridized nucleic acid. While it is preferred that the proportionality be relatively strict (*e.g.*, a doubling in transcription rate results in a doubling in mRNA transcript in the sample nucleic acid pool and a doubling in hybridization signal), one of skill will appreciate that the proportionality can be more relaxed and even non-linear. Thus, for example, an assay where a 5 fold difference in concentration of the target mRNA results in a 3 to 6 fold difference in hybridization intensity is sufficient for most purposes. Where more precise quantification is required appropriate controls can be run to correct for variations introduced in sample preparation and hybridization as described herein. In addition, serial dilutions of "standard" target mRNAs can be used to prepare calibration curves according to methods well known to those of skill in the art. Of course, where simple detection of the presence or absence of a transcript is desired, no elaborate control or calibration is required.

In the simplest embodiment, such a nucleic acid sample is the total mRNA isolated from a biological sample. The term "biological sample", as used herein, refers to a sample obtained from an organism or from components (*e.g.*, cells) of an organism. The sample may be of any biological tissue or fluid. Frequently the sample will be a "clinical sample" which is a sample derived from a patient. Such samples include, but

are not limited to, sputum, blood, blood cells (e.g., white cells), tissue or fine needle biopsy samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom. Biological samples may also include sections of tissues such as frozen sections taken for histological purposes.

The nucleic acid (either genomic DNA or mRNA) may be isolated from the sample according to any of a number of methods well known to those of skill in the art. One of skill will appreciate that where alterations in the copy number of a gene are to be detected genomic DNA is preferably isolated. Conversely, where expression levels of a gene or genes are to be detected, preferably RNA (mRNA) is isolated.

Methods of isolating total mRNA are well known to those of skill in the art. For example, methods of isolation and purification of nucleic acids are described in detail in Chapter 3 of *Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I. Theory and Nucleic Acid Preparation*, P. Tijssen, ed. Elsevier, N.Y. (1993) and Chapter 3 of *Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I. Theory and Nucleic Acid Preparation*, P. Tijssen, ed. Elsevier, N.Y. (1993)).

In a preferred embodiment, the total nucleic acid is isolated from a given sample using, for example, an acid guanidinium-phenol-chloroform extraction method and polyA$^+$ mRNA is isolated by oligo dT column chromatography or by using (dT)n magnetic beads (*see, e.g.*, Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual* (2nd ed.), Vols. 1-3, Cold Spring Harbor Laboratory, (1989), or *Current Protocols in Molecular Biology*, F. Ausubel *et al.*, ed. Greene Publishing and Wiley-Interscience, New York (1987)).

Frequently, it is desirable to amplify the nucleic acid sample prior to hybridization. One of skill in the art will appreciate that whatever amplification method is used, if a quantitative result is desired, care must be taken to use a method that maintains or controls for the relative frequencies of the amplified nucleic acids.

Methods of "quantitative" amplification are well known to those of skill in the art. For example, quantitative PCR involves simultaneously co-amplifying a known quantity of a control sequence using the same primers. This provides an internal standard that may be used to calibrate the PCR reaction. The high density array may

then include probes specific to the internal standard for quantification of the amplified nucleic acid.

One preferred internal standard is a synthetic AW106 cRNA. The AW106 cRNA is combined with RNA isolated from the sample according to standard techniques known to those of skill in the art. The RNA is then reverse transcribed using a reverse transcriptase to provide copy DNA. The cDNA sequences are then amplified (*e.g.*, by PCR) using labeled primers. The amplification products are separated, typically by electrophoresis, and the amount of radioactivity (proportional to the amount of amplified product) is determined. The amount of mRNA in the sample is then calculated by comparison with the signal produced by the known AW106 RNA standard. Detailed protocols for quantitative PCR are provided in *PCR Protocols, A Guide to Methods and Applications*, Innis *et al.*, Academic Press, Inc. N.Y., (1990).

Other suitable amplification methods include, but are not limited to polymerase chain reaction (PCR) (Innis, *et al., PCR Protocols. A guide to Methods and Application*. Academic Press, Inc. San Diego, (1990)), ligase chain reaction (LCR) (see Wu and Wallace, *Genomics*, 4: 560 (1989), Landegren, *et al., Science*, 241: 1077 (1988) and Barringer, *et al., Gene*, 89: 117 (1990), transcription amplification (Kwoh, et al., *Proc. Natl. Acad. Sci. USA*, 86: 1173 (1989)), and self-sustained sequence replication (Guatelli, *et al., Proc. Nat. Acad. Sci. USA*, 87: 1874 (1990)).

In a particularly preferred embodiment, the sample mRNA is reverse transcribed with a reverse transcriptase and a primer consisting of oligo dT and a sequence encoding the phage T7 promoter to provide single stranded DNA template. The second DNA strand is polymerized using a DNA polymerase. After synthesis of double-stranded cDNA, T7 RNA polymerase is added and RNA is transcribed from the cDNA template. Successive rounds of transcription from each single cDNA template results in amplified RNA. Methods of *in vitro* polymerization are well known to those of skill in the art (*see, e.g.*, Sambrook, *supra.*) and this particular method is described in detail by Van Gelder, *et al., Proc. Natl. Acad. Sci. USA*, 87: 1663-1667 (1990) who demonstrate that *in vitro* amplification according to this method preserves the relative frequencies of the various RNA transcripts. Moreover, Eberwine *et al. Proc. Natl. Acad. Sci. USA*, 89: 3010-3014 provide a protocol that uses two rounds of amplification

via *in vitro* transcription to achieve greater than 10^6 fold amplification of the original starting material thereby permitting expression monitoring even where biological samples are limited.

It will be appreciated by one of skill in the art that the direct transcription method described above provides an antisense (aRNA) pool. Where antisense RNA is used as the target nucleic acid, the oligonucleotide probes provided in the array are chosen to be complementary to subsequences of the antisense nucleic acids. Conversely, where the target nucleic acid pool is a pool of sense nucleic acids, the oligonucleotide probes are selected to be complementary to subsequences of the sense nucleic acids. Finally, where the nucleic acid pool is double stranded, the probes may be of either sense as the target nucleic acids include both sense and antisense strands.

The protocols cited above include methods of generating pools of either sense or antisense nucleic acids. Indeed, one approach can be used to generate either sense or antisense nucleic acids as desired. For example, the cDNA can be directionally cloned into a vector (*e.g.*, Stratagene's p Bluscript II KS (+) phagemid) such that it is flanked by the T3 and T7 promoters. *In vitro* transcription with the T3 polymerase will produce RNA of one sense (the sense depending on the orientation of the insert), while *in vitro* transcription with the T7 polymerase will produce RNA having the opposite sense. Other suitable cloning systems include phage lambda vectors designed for Cre-*loxP* plasmid subcloning (*see e.g.*, Palazzolo *et al.*, *Gene*, 88: 25-36 (1990)).

In a particularly preferred embodiment, a high activity RNA polymerase (*e.g.* about 2500 units/μL for T7, available from Epicentre Technologies) is used.

B) Labeling nucleic acids.

In a preferred embodiment, the hybridized nucleic acids are detected by detecting one or more labels attached to the sample nucleic acids. The labels may be incorporated by any of a number of means well known to those of skill in the art. However, in a preferred embodiment, the label is simultaneously incorporated during the amplification step in the preparation of the sample nucleic acids. Thus, for example, polymerase chain reaction (PCR) with labeled primers or labeled nucleotides will provide a labeled amplification product. In a preferred embodiment, transcription

amplification, as described above, using a labeled nucleotide (*e.g.* fluorescein-labeled UTP and/or CTP) incorporates a label into the transcribed nucleic acids.

Alternatively, a label may be added directly to the original nucleic acid sample (*e.g.*, mRNA, polyA mRNA, cDNA, *etc.*) or to the amplification product after the amplification is completed. Means of attaching labels to nucleic acids are well known to those of skill in the art and include, for example nick translation or end-labeling (*e.g.* with a labeled RNA) by kinasing of the nucleic acid and subsequent attachment (ligation) of a nucleic acid linker joining the sample nucleic acid to a label (*e.g.*, a fluorophore).

Detectable labels suitable for use in the present invention include any composition detectable by spectroscopic, photochemical, biochemical, immunochemical, electrical, optical or chemical means. Useful labels in the present invention include biotin for staining with labeled streptavidin conjugate, magnetic beads (*e.g.*, Dynabeads™), fluorescent dyes (*e.g.*, fluorescein, texas red, rhodamine, green fluorescent protein, and the like), radiolabels (*e.g.*, ^{3}H, ^{125}I, ^{35}S, ^{14}C, or ^{32}P), enzymes (*e.g.*, horse radish peroxidase, alkaline phosphatase and others commonly used in an ELISA), and colorimetric labels such as colloidal gold or colored glass or plastic (*e.g.*, polystyrene, polypropylene, latex, etc.) beads. Patents teaching the use of such labels include U.S. Patent Nos. 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149; and 4,366,241.

Means of detecting such labels are well known to those of skill in the art. Thus, for example, radiolabels may be detected using photographic film or scintillation counters, fluorescent markers may be detected using a photodetector to detect emitted light. Enzymatic labels are typically detected by providing the enzyme with a substrate and detecting the reaction product produced by the action of the enzyme on the substrate, and colorimetric labels are detected by simply visualizing the colored label.

The label may be added to the target (sample) nucleic acid(s) prior to, or after the hybridization. So called "direct labels" are detectable labels that are directly attached to or incorporated into the target (sample) nucleic acid prior to hybridization. In contrast, so called "indirect labels" are joined to the hybrid duplex after hybridization.

Often, the indirect label is attached to a binding moiety that has been attached to the target nucleic acid prior to the hybridization. Thus, for example, the target nucleic acid may be biotinylated before the hybridization. After hybridization, an aviden-conjugated fluorophore will bind the biotin bearing hybrid duplexes providing a label that is easily detected. For a detailed review of methods of labeling nucleic acids and detecting labeled hybridized nucleic acids see *Laboratory Techniques in Biochemistry and Molecular Biology, Vol. 24: Hybridization With Nucleic Acid Probes*, P. Tijssen, ed. Elsevier, N.Y., (1993)).

Fluorescent labels are preferred and easily added during an *in vitro* transcription reaction. In a preferred embodiment, fluorescein labeled UTP and CTP are incorporated into the RNA produced in an *in vitro* transcription reaction as described above.

C) Modifying sample to improve signal/noise ratio.

The nucleic acid sample may be modified prior to hybridization to the high density probe array in order to reduce sample complexity thereby decreasing background signal and improving sensitivity of the measurement. In one embodiment, complexity reduction is achieved by selective degradation of background mRNA. This is accomplished by hybridizing the sample mRNA (*e.g.*, polyA$^+$ RNA) with a pool of DNA oligonucleotides that hybridize specifically with the regions to which the probes in the array specifically hybridize. In a preferred embodiment, the pool of oligonucleotides consists of the same probe oligonucleotides as found on the high density array.

The pool of oligonucleotides hybridizes to the sample mRNA forming a number of double stranded (hybrid duplex) nucleic acids. The hybridized sample is then treated with RNase A, a nuclease that specifically digests single stranded RNA. The RNase A is then inhibited, using a protease and/or commercially available RNase inhibitors, and the double stranded nucleic acids are then separated from the digested single stranded RNA. This separation may be accomplished in a number of ways well known to those of skill in the art including, but not limited to, electrophoresis, and gradient centrifugation. However, in a preferred embodiment, the pool of DNA oligonucleotides is provided attached to beads forming thereby a nucleic acid affinity

column. After digestion with the RNase A, the hybridized DNA is removed simply by denaturing (*e.g.*, by adding heat or increasing salt) the hybrid duplexes and washing the previously hybridized mRNA off in an elution buffer.

The undigested mRNA fragments which will be hybridized to the probes in the high density array are then preferably end-labeled with a fluorophore attached to an RNA linker using an RNA ligase. This procedure produces a labeled sample RNA pool in which the nucleic acids that do not correspond to probes in the array are eliminated and thus unavailable to contribute to a background signal.

Another method of reducing sample complexity involves hybridizing the mRNA with deoxyoligonucleotides that hybridize to regions that border on either size the regions to which the high density array probes are directed. Treatment with RNAse H selectively digests the double stranded (hybrid duplexes) leaving a pool of single-stranded mRNA corresponding to the short regions (*e.g.*, 20 mer) that were formerly bounded by the deoxyoligonucleotide probes and which correspond to the targets of the high density array probes and longer mRNA sequences that correspond to regions between the targets of the probes of the high density array. The short RNA fragments are then separated from the long fragments (*e.g.*, by electrophoresis), labeled if necessary as described above, and then are ready for hybridization with the high density probe array.

In a third approach, sample complexity reduction involves the selective removal of particular (preselected) mRNA messages. In particular, highly expressed mRNA messages that are not specifically probed by the probes in the high density array are preferably removed. This approach involves hybridizing the polyA$^+$ mRNA with an oligonucleotide probe that specifically hybridizes to the preselected message close to the 3' (poly A) end. The probe may be selected to provide high specificity and low cross reactivity. Treatment of the hybridized message/probe complex with RNase H digests the double stranded region effectively removing the polyA$^+$ tail from the rest of the message. The sample is then treated with methods that specifically retain or amplify polyA$^+$ RNA (*e.g.*, an oligo dT column or (dT)n magnetic beads). Such methods will not retain or amplify the selected message(s) as they are no longer associated with a

polyA$^+$ tail. These highly expressed messages are effectively removed from the sample providing a sample that has reduced background mRNA.

IV. Hybridization Array Design.

A) Probe composition.

One of skill in the art will appreciate that an enormous number of array designs are suitable for the practice of this invention. The high density array will typically include a number of probes that specifically hybridize to the nucleic acid(s) expression of which is to be detected. In addition, in a preferred embodiment, the array will include one or more control probes.

1) Test probes.

In its simplest embodiment, the high density array includes "test probes". These are oligonucleotides that range from about 5 to about 45 or 5 to about 50 nucleotides, more preferably from about 10 to about 40 nucleotides and most preferably from about 15 to about 40 nucleotides in length. In other particularly preferred embodiments the probes are 20 or 25 nucleotides in length. These oligonucleotide probes have sequences complementary to particular subsequences of the genes whose expression they are designed to detect. Thus, the test probes are capable of specifically hybridizing to the target nucleic acid they are to detect.

In addition to test probes that bind the target nucleic acid(s) of interest, the high density array can contain a number of control probes. The control probes fall into three categories referred to herein as 1) Normalization controls; 2) Expression level controls; and 3) Mismatch controls.

2) Normalization controls.

Normalization controls are oligonucleotide probes that are perfectly complementary to labeled reference oligonucleotides that are added to the nucleic acid sample. The signals obtained from the normalization controls after hybridization provide a control for variations in hybridization conditions, label intensity, "reading" efficiency and other factors that may cause the signal of a perfect hybridization to vary between arrays. In a preferred embodiment, signals (*e.g.*, fluorescence intensity) read

from all other probes in the array are divided by the signal (*e.g.*, fluorescence intensity) from the control probes thereby normalizing the measurements.

Virtually any probe may serve as a normalization control. However, it is recognized that hybridization efficiency varies with base composition and probe length. Preferred normalization probes are selected to reflect the average length of the other probes present in the array, however, they can be selected to cover a range of lengths. The normalization control(s) can also be selected to reflect the (average) base composition of the other probes in the array, however in a preferred embodiment, only one or a few normalization probes are used and they are selected such that they hybridize well (*i.e.* no secondary structure) and do not match any target-specific probes.

Normalization probes can be localized at any position in the array or at multiple positions throughout the array to control for spatial variation in hybridization efficiently. In a preferred embodiment, the normalization controls are located at the corners or edges of the array as well as in the middle.

3) Expression level controls.

Expression level controls are probes that hybridize specifically with constitutively expressed genes in the biological sample. Expression level controls are designed to control for the overall health and metabolic activity of a cell. Examination of the covariance of an expression level control with the expression level of the target nucleic acid indicates whether measured changes or variations in expression level of a gene is due to changes in transcription rate of that gene or to general variations in health of the cell. Thus, for example, when a cell is in poor health or lacking a critical metabolite the expression levels of both an active target gene and a constitutively expressed gene are expected to decrease. The converse is also true. Thus where the expression levels of both an expression level control and the target gene appear to both decrease or to both increase, the change may be attributed to changes in the metabolic activity of the cell as a whole, not to differential expression of the target gene in question. Conversely, where the expression levels of the target gene and the expression level control do not covary, the variation in the expression level of the target gene is

attributed to differences in regulation of that gene and not to overall variations in the metabolic activity of the cell.

Virtually any constitutively expressed gene provides a suitable target for expression level controls. Typically expression level control probes have sequences complementary to subsequences of constitutively expressed "housekeeping genes" including, but not limited to the ß-actin gene, the transferrin receptor gene, the GAPDH gene, and the like.

4) Mismatch controls.

Mismatch controls may also be provided for the probes to the target genes, for expression level controls or for normalization controls. Mismatch controls are oligonucleotide probes identical to their corresponding test or control probes except for the presence of one or more mismatched bases. A mismatched base is a base selected so that it is not complementary to the corresponding base in the target sequence to which the probe would otherwise specifically hybridize. One or more mismatches are selected such that under appropriate hybridization conditions (e.g. stringent conditions) the test or control probe would be expected to hybridize with its target sequence, but the mismatch probe would not hybridize (or would hybridize to a significantly lesser extent). Preferred mismatch probes contain a central mismatch. Thus, for example, where a probe is a 20 mer, a corresponding mismatch probe will have the identical sequence except for a single base mismatch (*e.g.*, substituting a G, a C or a T for an A) at any of positions 6 through 14 (the central mismatch).

Mismatch probes thus provide a control for non-specific binding or cross-hybridization to a nucleic acid in the sample other than the target to which the probe is directed. Mismatch probes thus indicate whether a hybridization is specific or not. For example, if the target is present the perfect match probes should be consistently brighter than the mismatch probes. In addition, if all central mismatches are present, the mismatch probes can be used to detect a mutation. Finally, it was also a discovery of the present invention that the difference in intensity between the perfect match and the mismatch probe (I(PM)-I(MM)) provides a good measure of the concentration of the hybridized material.

5) Sample preparation/amplification controls.

The high density array may also include sample preparation/amplification control probes. These are probes that are complementary to subsequences of control genes selected because they do not normally occur in the nucleic acids of the particular biological sample being assayed. Suitable sample preparation/amplification control probes include, for example, probes to bacterial genes (*e.g.*, Bio B) where the sample in question is a biological from a eukaryote.

The RNA sample is then spiked with a known amount of the nucleic acid to which the sample preparation/amplification control probe is directed before processing. Quantification of the hybridization of the sample preparation/amplification control probe then provides a measure of alteration in the abundance of the nucleic acids caused by processing steps (*e.g.* PCR, reverse transcription, *in vitro* transcription, *etc.*).

B) Probe Selection and Optimization.

In a preferred embodiment, oligonucleotide probes in the high density array are selected to bind specifically to the nucleic acid target to which they are directed with minimal non-specific binding or cross-hybridization under the particular hybridization conditions utilized. Because the high density arrays of this invention can contain in excess of 1,000,000 different probes, it is possible to provide every probe of a characteristic length that binds to a particular nucleic acid sequence. Thus, for example, the high density array can contain every possible 20 mer sequence complementary to an IL-2 mRNA.

There, however, may exist 20 mer subsequences that are not unique to the IL-2 mRNA. Probes directed to these subsequences are expected to cross hybridize with occurrences of their complementary sequence in other regions of the sample genome. Similarly, other probes simply may not hybridize effectively under the hybridization conditions (*e.g.*, due to secondary structure, or interactions with the substrate or other probes). Thus, in a preferred embodiment, the probes that show such poor specificity or hybridization efficiency are identified and may not be included either in the high density array itself (*e.g.*, during fabrication of the array) or in the post-hybridization data analysis.

In addition, in a preferred embodiment, expression monitoring arrays are used to identify the presence and expression (transcription) level of genes which are several hundred base pairs long. For most applications it would be useful to identify the presence, absence, or expression level of several thousand to one hundred thousand genes. Because the number of oligonucleotides per array is limited in a preferred embodiment, it is desired to include only a limited set of probes specific to each gene whose expression is to be detected.

It is a discovery of this invention that probes as short as 15, 20, or 25 nucleotides are sufficient to hybridize to a subsequence of a gene and that, for most genes, there is a set of probes that performs well across a wide range of target nucleic acid concentrations. In a preferred embodiment, it is desirable to choose a preferred or "optimum" subset of probes for each gene before synthesizing the high density array.

1) Hybridization and Cross-Hybridization Data.

Thus, in one embodiment, this invention provides for a method of optimizing a probe set for detection of a particular gene. Generally, this method involves providing a high density array containing a multiplicity of probes of one or more particular length(s) that are complementary to subsequences of the mRNA transcribed by the target gene. In one embodiment the high density array may contain every probe of a particular length that is complementary to a particular mRNA. The probes of the high density array are then hybridized with their target nucleic acid alone and then hybridized with a high complexity, high concentration nucleic acid sample that does not contain the targets complementary to the probes. Thus, for example, where the target nucleic acid is an RNA, the probes are first hybridized with their target nucleic acid alone and then hybridized with RNA made from a cDNA library (*e.g.*, reverse transcribed polyA$^+$ mRNA) where the sense of the hybridized RNA is opposite that of the target nucleic acid (to insure that the high complexity sample does not contain targets for the probes). Those probes that show a strong hybridization signal with their target and little or no cross-hybridization with the high complexity sample are preferred probes for use in the high density arrays of this invention.

The high density array may additionally contain mismatch controls for each of the probes to be tested. In a preferred embodiment, the mismatch controls contain a central mismatch. Where both the mismatch control and the target probe show high levels of hybridization (*e.g.*, the hybridization to the mismatch is nearly equal to or greater than the hybridization to the corresponding test probe), the test probe is preferably not used in the high density array.

In a particularly preferred embodiment, optimal probes are selected according to the following method: First, as indicated above, an array is provided containing a multiplicity of oligonucleotide probes complementary to subsequences of the target nucleic acid. The oligonucleotide probes may be of a single length or may span a variety of lengths ranging from 5 to 50 nucleotides. The high density array may contain every probe of a particular length that is complementary to a particular mRNA or may contain probes selected from various regions of particular mRNAs. For each target-specific probe the array also contains a mismatch control probe; preferably a central mismatch control probe.

The oligonucleotide array is hybridized to a sample containing target nucleic acids having subsequences complementary to the oligonucleotide probes and the difference in hybridization intensity between each probe and its mismatch control is determined. Only those probes where the difference between the probe and its mismatch control exceeds a threshold hybridization intensity (*e.g.* preferably greater than 10% of the background signal intensity, more preferably greater than 20% of the background signal intensity and most preferably greater than 50% of the background signal intensity) are selected. Thus, only probes that show a strong signal compared to their mismatch control are selected.

The probe optimization procedure can optionally include a second round of selection. In this selection, the oligonucleotide probe array is hybridized with a nucleic acid sample that is not expected to contain sequences complementary to the probes. Thus, for example, where the probes are complementary to the RNA sense strand a sample of antisense RNA is provided. Of course, other samples could be provided such as samples from organisms or cell lines known to be lacking a particular gene, or known for not expressing a particular gene.

Only those probes where both the probe and its mismatch control show hybridization intensities below a threshold value (*e.g.* less than about 5 times the background signal intensity, preferably equal to or less than about 2 times the background signal intensity, more preferably equal to or less than about 1 times the background signal intensity, and most preferably equal or less than about half background signal intensity) are selected. In this way probes that show minimal non-specific binding are selected. Finally, in a preferred embodiment, the n probes (where n is the number of probes desired for each target gene) that pass both selection criteria and have the highest hybridization intensity for each target gene are selected for incorporation into the array, or where already present in the array, for subsequent data analysis. Of course, one of skill in the art, will appreciate that either selection criterion could be used alone for selection of probes.

2) Heuristic rules.

Using the hybridization and cross-hybridization data obtained as described above, graphs can be made of hybridization and cross-hybridization intensities versus various probe properties *e.g.*, number of As, number of Cs in a window of 8 bases, palindomic strength, *etc.* The graphs can then be examined for correlations between those properties and the hybridization or cross-hybridization intensities. Thresholds can be set beyond which it looks like hybridization is always poor or cross hybridization is always very strong. If any probe fails one of the criteria, it is rejected from the set of probes and therefore, not placed on the chip. This will be called the heuristic rules method.

One set of rules developed for 20 mer probes in this manner is the following:

Hybridization rules:

1) Number of As is less than 9.

2) Number of Ts is less than 10 and greater than 0.

3) Maximum run of As, Gs, or Ts is less than 4 bases in a row.

4) Maximum run of any 2 bases is less than 11 bases.

5) Palindrome score is less than 6.

6) Clumping score is less than 6.

7) Number of As + Number of Ts is less than 14

8) Number of As + number of Gs is less than 15

With respect to rule number 4, requiring the maximum run of any two bases to be less than 11 bases guarantees that at least three different bases occur within any 12 consecutive nucleotides. A palindrome score is the maximum number of complementary bases if the oligonucleotide is folded over at a point that maximizes self complementarity. Thus, for example a 20 mer that is perfectly self-complementary would have a palindrome score of 10. A clumping score is the maximum number of three-mers of identical bases in a given sequence. Thus, for example, a run of 5 identical bases will produce a clumping score of 3 (bases 1-3, bases 2-4, and bases 3-5).

If any probe failed one of these criteria (1-8), the probe was not a member of the subset of probes placed on the chip. For example, if a hypothetical probe was 5'-AGCTTTTTTCATGCATCTAT-3' the probe would not be synthesized on the chip because it has a run of four or more bases (*i.e.*, run of six).

The cross hybridization rules developed for 20 mers were as follows:

1) Number of Cs is less than 8;

2) Number of Cs in any window of 8 bases is less than 4.

Thus, if any probe failed any of either the hybridization ruses (1-8) or the cross-hybridization rules (1-2), the probe was not a member of the subset of probes placed on the chip. These rules eliminated many of the probes that cross hybridized strongly or exhibited low hybridization, and performed moderate job of eliminating weakly hybridizing probes.

These heuristic rules may be implemented by hand calculations, or alternatively, they may be implemented in software as is discussed below in Section IV.B.7.

3) Neural net.

In another embodiment, a neural net can be trained to predict the hybridization and cross-hybridization intensities based on the sequence of the probe or on other probe properties. The neural net can then be used to pick an arbitrary number

of the "best" probes. One such neural net was developed for selecting 20-mer probes. This neural net was produced a moderate (0.7) correlation between predicted intensity and measured intensity, with a better model for cross hybridization than hybridization. Details of this neural net are provided in Example 6.

5

4) ANOVA Model

An analysis of variance (ANOVA) model may be built to model the intensities based on positions of consecutive base pairs. This is based on the theory that the melting energy is based on stacking energies of consecutive bases. The annova
10 model was used to find correlation between the a probe sequence and the hybridization and cross-hybridization intensities. The inputs were probe sequences broken down into consecutive base pairs. One model was made to predict hybridization, another was made to predict cross hybridization. The output was the hybridization or crosshybridization intensity.

15 There were 304 (19 * 16) possible inputs, consisting of the 14 possible two base combinations, and the 19 positions that those combinations could be found in. For example, the sequence aggctga... has "ag" in the first position, "gg" in the second position, "gc" in the third, "ct" in the fourth and so on.

The resulting model assigned a component of the output intensity to each
20 of the possible inputs, so to estimate the intensity for a given sequence one simply adds the intensities for each of it's 19 components.

5) Pruning (removal) of similar probes.

One of the causes of poor signals in expression chips is that genes other than
25 the ones being monitored have sequences which are very similar to parts of the sequences which are being monitored. The easiest way to solve this is to remove probes which are similar to more than one gene. Thus, in a preferred embodiment, it is desirable to remove (prune) probes that hybridize to transcription products of more than one gene.

The simplest pruning method is to line up a proposed probe with all known
30 genes for the organism being monitored, then count the number of matching bases. For example, given a probe to gene 1 of an organism and gene 2 of an organism as follows:

```
probe from gene 1:  aagcgcgatcgattatgctc
                    |      |||||||
gene 2:             atctcggatcgatcggataagcgcgatcgattatgctcggcga
```

has 8 matching bases in this alignment, but 20 matching bases in the following alignment:

```
probe from gene 1:                    aagcgcgatcgattatgctc
                                      ||||||||||||||||||||
gene 2:             atctcggatcgatcggataagcgcgatcgattatgctcggcga
```

More complicated algorithms also exist, which allow the detection of insertion or deletion mismatches. Such sequence alignment algorithms are well known to those of skill in the art and include, but are not limited to BLAST, or FASTA, or other gene matching programs such as those described above in the definitions section.

In another variant, where an organism has many different genes which are very similar, it is difficult to make a probe set that measures the concentration only one of those very similar genes. One can then prune out any probes which are dissimilar, and make the probe set a probe set for that family of genes.

(f) Synthesis cycle pruning.

The cost of producing masks for a chip is approximately linearly related to the number of synthesis cycles. In a normal set of genes the distribution of the number of cycles any probe takes to build approximates a Gausian distribution. Because of this the mask cost can normally be reduced by 15% by throwing out about 3 percent of the probes. In a preferred embodiment, synthesis cycle pruning simply involves eliminating (not including) those probes those probes that require a greater number of synthesis cycles than the maximum number of synthesis cycles selected for preparation of the particular subject high density oligonucleotide array. Since the typical synthesis of probes follows a regular pattern of bases put down (acgtacgtacgt...) counting the number of synthesis steps needed to build a probe is easy. The listing shown in Table 1 povides typical code for counting the number of synthesis cycles a probe will need.

Table 1. Typical code for counting synthesis cycles required for the chemical synthesis of a probe.

```
static char base[] = "acgt";
//                      a b c d e f g h i j k l m n o p q r s t u v w x y z
static short index[] = { 0, 0, 1, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0};

short lookupIndex( char aBase ){
        if( isupper( aBase ) || !isalpha( aBase) ){
                errorHwnd( "illegal base");
                return -1;
        }
        if( strchr( base, aBase ) == NULL ){
                errorHwnd( "non-dna base");
                return 0;
        }
  return index[ aBase - 'a'];
}

static short calculateMinNumberOfSynthesisStepsForComplement( char local * buffer ){
        short i, last, current, cycles = 1;
        char buffer1[40];
        for( i =3D 0; buffer[i] != 0; i++ ){
                switch( tolower(buffer[i]) ){
                case 'a':  buffer1[i] = 't';break;
                case 'c':  buffer1[i] = 'g';break;
                case 'g':  buffer1[i] = 'c';break;
                case 't':  buffer1[i] = 'a';break;
                }
        }
        buffer1[i] = 0;
        if( buffer1[0] == 0 ) return 0;
        last = lookupIndex( buffer1[0] );
        for( i = 1; buffer1[i] != 0; i++ ){
                current = lookupIndex( buffer1[i] );
                if( current <= last ) cycles++;
                last = current;
        }
        return (short)((cycles -1) * 4 + current +1);
}
```

7) Combination of Selection methods.

The heuristic rules, neural net and annova model provide ways of pruning or reducing the number of probes for monitoring the expression of genes. As these methods

do not necessarily produce the same results, or produce entirely independent results, it may be advantageous to combine the methods. For example, probes may be pruned or reduced if more than one method (*e.g.*, two out of three) indicate the probe will not likely produce good results. Then, synthesis cycle pruning may be performed to reduce costs.

Fig. 11 shows the flow of a process of increasing the number of probes for monitoring the expression of genes after the number of probes has been reduced or pruned. In one embodiment, a user is able to specify the number of nucleic acid probes that should be placed on the chip to monitor the expression of each gene. As discussed above, it is advantageous to reduce probes that will not likely produce good results; however, the number of probes may be reduced to substantially less than the desired number of probes.

At step 402, the number of probes for monitoring multiple genes is reduced by the heuristic rules method, neural net, annova model, synthesis cycle pruning, or any other method, or combination of methods. A gene is selected at step 404.

A determination is made whether the remaining probes for monitoring the selected gene number greater than 80% (which may be varied or user defined) of the desired number of probes. If yes, the computer system proceeds to the next gene at step 408 which will generally return to step 404.

If the remaining probes for monitoring the selected gene do not number greater than 80% of the desired number of probes, a determination is made whether the remaining probes for monitoring the selected gene number greater than 40% (which may be varied or user defined) of the desired number of probes. If yes, an "i" is appended to the end of the gene name to indicate that after pruning, the probes were incomplete at step 412.

At step 414, the number of probes is increased by loosening the constraints that rejected probes. For example, the thresholds in the heuristic rules may be increased by 1. Therefore, if previously probes were rejected if they had four As in a row, the rule may be loosened to five As in a row.

A determination is then made whether the remaining probes for monitoring the selected gene number greater than 80% of the desired number of probes at step 416. If yes, an "r" is appended to the end of the gene name at step 412 to indicate that the rules were loosened to generate the number of synthesized probes for that gene.

At step 420, a check is made to see if the probes for monitoring the selected gene only conflict with one or two other genes. If yes, the full set of probes complementary to the gene (or target sequence) are taken and pruned so that the probes remaining are exactly complementary to the selected gene exclusively at step 422.

A determination is then made whether the remaining probes for monitoring the selected gene number greater than 80% of the desired number of probes at step 424. If yes, an "s" is appended to the end of the gene name at step 426 to indicate that the only a few genes were similar to the selected gene.

At step 428, the probes for monitoring the selected gene are not reduced by conflicts at all. A determination is then made whether the remaining probes for monitoring the selected gene number greater than 80% of the desired number of probes at step 430. If yes, an "f" is appended to the end of the gene name at step 432 to indicate that the probes include the whole family of probes perfectly complementary to the gene.

If there are still not 80% of the desired number of probes, an error is reported at step 434. Any number of error handling procedures may be undertaken. For example, an error message may be generated for the user and the probes for the gene may not be stored. Alternatively, the user may be prompted to enter a new desired number of probes.

V. Synthesis of High Density Arrays

Methods of forming high density arrays of oligonucleotides, peptides and other polymer sequences with a minimal number of synthetic steps are known. The oligonucleotide analogue array can be synthesized on a solid substrate by a variety of methods, including, but not limited to, light-directed chemical coupling, and mechanically directed coupling. See Pirrung *et al.*, U.S. Patent No. 5,143,854 (see also PCT Application No. WO 90/15070) and Fodor *et al.*, PCT Publication Nos. WO 92/10092 and WO 93/09668 which disclose methods of forming vast arrays of peptides, oligonucleotides and other molecules using, for example, light-directed synthesis techniques. See also, Fodor *et al.*, *Science*, 251, 767-77 (1991). These procedures for synthesis of polymer arrays are now referred to as VLSIPS™ procedures. Using the VLSIPS™ approach, one heterogenous array of polymers is converted, through

simultaneous coupling at a number of reaction sites, into a different heterogenous array. See, U.S. Application Serial Nos. 07/796,243 and 07/980,523.

The development of VLSIPS™ technology as described in the above-noted U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, is considered pioneering technology in the fields of combinatorial synthesis and screening of combinatorial libraries. More recently, patent application Serial No. 08/082,937, filed June 25, 1993 describes methods for making arrays of oligonucleotide probes that can be used to check or determine a partial or complete sequence of a target nucleic acid and to detect the presence of a nucleic acid containing a specific oligonucleotide sequence.

In brief, the light-directed combinatorial synthesis of oligonucleotide arrays on a glass surface proceeds using automated phosphoramidite chemistry and chip masking techniques. In one specific implementation, a glass surface is derivatized with a silane reagent containing a functional group, *e.g.*, a hydroxyl or amine group blocked by a photolabile protecting group. Photolysis through a photolithogaphic mask is used selectively to expose functional groups which are then ready to react with incoming 5'-photoprotected nucleoside phosphoramidites. The phosphoramidites react only with those sites which are illuminated (and thus exposed by removal of the photolabile blocking group). Thus, the phosphoramidites only add to those areas selectively exposed from the preceding step. These steps are repeated until the desired array of sequences have been synthesized on the solid surface. Combinatorial synthesis of different oligonucleotide analogues at different locations on the array is determined by the pattern of illumination during synthesis and the order of addition of coupling reagents.

In the event that an oligonucleotide analogue with a polyamide backbone is used in the VLSIPS™ procedure, it is generally inappropriate to use phosphoramidite chemistry to perform the synthetic steps, since the monomers do not attach to one another via a phosphate linkage. Instead, peptide synthetic methods are substituted. See, *e.g.*, Pirrung *et al.* U.S. Pat. No. 5,143,854.

Peptide nucleic acids are commercially available from, *e.g.*, Biosearch, Inc. (Bedford, MA) which comprise a polyamide backbone and the bases found in naturally occurring nucleosides. Peptide nucleic acids are capable of binding to nucleic

acids with high specificity, and are considered "oligonucleotide analogues" for purposes of this disclosure.

In addition to the foregoing, additional methods which can be used to generate an array of oligonucleotides on a single substrate are described in co-pending Applications Ser. No. 07/980,523, filed November 20, 1992, and 07/796,243, filed November 22, 1991 and in PCT Publication No. WO 93/09668. In the methods disclosed in these applications, reagents are delivered to the substrate by either (1) flowing within a channel defined on predefined regions or (2) "spotting" on predefined regions. However, other approaches, as well as combinations of spotting and flowing, may be employed. In each instance, certain activated regions of the substrate are mechanically separated from other regions when the monomer solutions are delivered to the various reaction sites.

A typical "flow channel" method applied to the compounds and libraries of the present invention can generally be described as follows. Diverse polymer sequences are synthesized at selected regions of a substrate or solid support by forming flow channels on a surface of the substrate through which appropriate reagents flow or in which appropriate reagents are placed. For example, assume a monomer "A" is to be bound to the substrate in a first group of selected regions. If necessary, all or part of the surface of the substrate in all or a part of the selected regions is activated for binding by, for example, flowing appropriate reagents through all or some of the channels, or by washing the entire substrate with appropriate reagents. After placement of a channel block on the surface of the substrate, a reagent having the monomer A flows through or is placed in all or some of the channel(s). The channels provide fluid contact to the first selected regions, thereby binding the monomer A on the substrate directly or indirectly (via a spacer) in the first selected regions.

Thereafter, a monomer B is coupled to second selected regions, some of which may be included among the first selected regions. The second selected regions will be in fluid contact with a second flow channel(s) through translation, rotation, or replacement of the channel block on the surface of the substrate; through opening or closing a selected valve; or through deposition of a layer of chemical or photoresist. If necessary, a step is performed for activating at least the second regions. Thereafter,

the monomer B is flowed through or placed in the second flow channel(s), binding monomer B at the second selected locations. In this particular example, the resulting sequences bound to the substrate at this stage of processing will be, for example, A, B, and AB. The process is repeated to form a vast array of sequences of desired length at known locations on the substrate.

After the substrate is activated, monomer A can be flowed through some of the channels, monomer B can be flowed through other channels, a monomer C can be flowed through still other channels, *etc.* In this manner, many or all of the reaction regions are reacted with a monomer before the channel block must be moved or the substrate must be washed and/or reactivated. By making use of many or all of the available reaction regions simultaneously, the number of washing and activation steps can be minimized.

One of skill in the art will recognize that there are alternative methods of forming channels or otherwise protecting a portion of the surface of the substrate. For example, according to some embodiments, a protective coating such as a hydrophilic or hydrophobic coating (depending upon the nature of the solvent) is utilized over portions of the substrate to be protected, sometimes in combination with materials that facilitate wetting by the reactant solution in other regions. In this manner, the flowing solutions are further prevented from passing outside of their designated flow paths.

The "spotting" methods of preparing compounds and libraries of the present invention can be implemented in much the same manner as the flow channel methods. For example, a monomer A can be delivered to and coupled with a first group of reaction regions which have been appropriately activated. Thereafter, a monomer B can be delivered to and reacted with a second group of activated reaction regions. Unlike the flow channel embodiments described above, reactants are delivered by directly depositing (rather than flowing) relatively small quantities of them in selected regions. In some steps, of course, the entire substrate surface can be sprayed or otherwise coated with a solution. In preferred embodiments, a dispenser moves from region to region, depositing only as much monomer as necessary at each stop. Typical dispensers include a micropipette to deliver the monomer solution to the substrate and a robotic system to control the position of the micropipette with respect to the substrate.

In other embodiments, the dispenser includes a series of tubes, a manifold, an array of pipettes, or the like so that various reagents can be delivered to the reaction regions simultaneously.

VI. Hybridization.

Nucleic acid hybridization simply involves providing a denatured probe and target nucleic acid under conditions where the probe and its complementary target can form stable hybrid duplexes through complementary base pairing. The nucleic acids that do not form hybrid duplexes are then washed away leaving the hybridized nucleic acids to be detected, typically through detection of an attached detectable label. It is generally recognized that nucleic acids are denatured by increasing the temperature or decreasing the salt concentration of the buffer containing the nucleic acids. Under low stringency conditions (*e.g.*, low temperature and/or high salt) hybrid duplexes (*e.g.*, DNA:DNA, RNA:RNA, or RNA:DNA) will form even where the annealed sequences are not perfectly complementary. Thus specificity of hybridization is reduced at lower stringency. Conversely, at higher stringency (*e.g.*, higher temperature or lower salt) successful hybridization requires fewer mismatches.

One of skill in the art will appreciate that hybridization conditions may be selected to provide any degree of stringency. In a preferred embodiment, hybridization is performed at low stringency in this case in 6X SSPE-T at 37°C (0.005% Triton X-100) to ensure hybridization and then subsequent washes are performed at higher stringency (*e.g.*, 1 X SSPE-T at 37°C) to eliminate mismatched hybrid duplexes. Successive washes may be performed at increasingly higher stringency (*e.g.*, down to as low as 0.25 X SSPE-T at 37°C to 50°C) until a desired level of hybridization specificity is obtained. Stringency can also be increased by addition of agents such as formamide. Hybridization specificity may be evaluated by comparison of hybridization to the test probes with hybridization to the various controls that can be present (*e.g.*, expression level control, normalization control, mismatch controls, *etc.*).

In general, there is a tradeoff between hybridization specificity (stringency) and signal intensity. Thus, in a preferred embodiment, the wash is performed at the highest stringency that produces consistent results and that provides a

signal intensity greater than approximately 10% of the background intensity. Thus, in a preferred embodiment, the hybridized array may be washed at successively higher stringency solutions and read between each wash. Analysis of the data sets thus produced will reveal a wash stringency above which the hybridization pattern is not appreciably altered and which provides adequate signal for the particular oligonucleotide probes of interest.

In a preferred embodiment, background signal is reduced by the use of a detergent (*e.g.*, C-TAB) or a blocking reagent (*e.g.*, sperm DNA, cot-1 DNA, etc.) during the hybridization to reduce non-specific binding. In a particularly preferred embodiment, the hybridization is performed in the presence of about 0.5 mg/ml DNA (*e.g.*, herring sperm DNA). The use of blocking agents in hybridization is well known to those of skill in the art (*see, e.g.*, Chapter 8 in P. Tijssen, *supra.*)

The stability of duplexes formed between RNAs or DNAs are generally in the order of RNA:RNA > RNA:DNA > DNA:DNA, in solution. Long probes have better duplex stability with a target, but poorer mismatch discrimination than shorter probes (mismatch discrimination refers to the measured hybridization signal ratio between a perfect match probe and a single base mismatch probe). Shorter probes (*e.g.*, 8-mers) discriminate mismatches very well, but the overall duplex stability is low.

Altering the thermal stability (T_m) of the duplex formed between the target and the probe using, *e.g.*, known oligonucleotide analogues allows for optimization of duplex stability and mismatch discrimination. One useful aspect of altering the T_m arises from the fact that adenine-thymine (A-T) duplexes have a lower T_m than guanine-cytosine (G-C) duplexes, due in part to the fact that the A-T duplexes have 2 hydrogen bonds per base-pair, while the G-C duplexes have 3 hydrogen bonds per base pair. In heterogeneous oligonucleotide arrays in which there is a non-uniform distribution of bases, it is not generally possible to optimize hybridization for each oligonucleotide probe simultaneously. Thus, in some embodiments, it is desirable to selectively destabilize G-C duplexes and/or to increase the stability of A-T duplexes. This can be accomplished, *e.g.*, by substituting guanine residues in the probes of an array which form G-C duplexes with hypoxanthine, or by substituting adenine residues in probes

which form A-T duplexes with 2,6 diaminopurine or by using the salt tetramethyl ammonium chloride (TMACl) in place of NaCl.

Altered duplex stability conferred by using oligonucleotide analogue probes can be ascertained by following, *e.g.*, fluorescence signal intensity of oligonucleotide analogue arrays hybridized with a target oligonucleotide over time. The data allow optimization of specific hybridization conditions at, *e.g.*, room temperature (for simplified diagnostic applications in the future).

Another way of verifying altered duplex stability is by following the signal intensity generated upon hybridization with time. Previous experiments using DNA targets and DNA chips have shown that signal intensity increases with time, and that the more stable duplexes generate higher signal intensities faster than less stable duplexes. The signals reach a plateau or "saturate" after a certain amount of time due to all of the binding sites becoming occupied. These data allow for optimization of hybridization, and determination of the best conditions at a specified temperature.

Methods of optimizing hybridization conditions are well known to those of skill in the art (*see, e.g., Laboratory Techniques in Biochemistry and Molecular Biology, Vol. 24: Hybridization With Nucleic Acid Probes*, P. Tijssen, ed. Elsevier, N.Y., (1993)).

VII. Signal Detection.

Means of detecting labeled target (sample) nucleic acids hybridized to the probes of the high density array are known to those of skill in the art. Thus, for example, where a colorimetric label is used, simple visualization of the label is sufficient. Where a radioactive labeled probe is used, detection of the radiation (*e.g* with photographic film or a solid state detector) is sufficient.

In a preferred embodiment, however, the target nucleic acids are labeled with a fluorescent label and the localization of the label on the probe array is accomplished with fluorescent microscopy. The hybridized array is excited with a light source at the excitation wavelength of the particular fluorescent label and the resulting fluorescence at the emission wavelength is detected. In a particularly preferred

embodiment, the excitation light source is a laser appropriate for the excitation of the fluorescent label.

The confocal microscope may be automated with a computer-controlled stage to automatically scan the entire high density array. Similarly, the microscope may be equipped with a phototransducer (*e.g.*, a photomultiplier, a solid state array, a ccd camera, *etc.*) attached to an automated data acquisition system to automatically record the fluorescence signal produced by hybridization to each oligonucleotide probe on the array. Such automated systems are described at length in U.S. Patent No: 5,143,854, PCT Application 20 92/10092, and copending U.S.S.N. 08/195,889 filed on February 10, 1994. Use of laser illumination in conjunction with automated confocal microscopy for signal detection permits detection at a resolution of better than about 100 μm, more preferably better than about 50 μm, and most preferably better than about 25 μm.

VIII. Signal Evaluation.

One of skill in the art will appreciate that methods for evaluating the hybridization results vary with the nature of the specific probe nucleic acids used as well as the controls provided. In the simplest embodiment, simple quantification of the fluorescence intensity for each probe is determined. This is accomplished simply by measuring probe signal strength at each location (representing a different probe) on the high density array (*e.g.*, where the label is a fluorescent label, detection of the amount of florescence (intensity) produced by a fixed excitation illumination at each location on the array). Comparison of the absolute intensities of an array hybridized to nucleic acids from a "test" sample with intensities produced by a "control" sample provides a measure of the relative expression of the nucleic acids that hybridize to each of the probes.

One of skill in the art, however, will appreciate that hybridization signals will vary in strength with efficiency of hybridization, the amount of label on the sample nucleic acid and the amount of the particular nucleic acid in the sample. Typically nucleic acids present at very low levels (*e.g.*, < 1pM) will show a very weak signal. At some low level of concentration, the signal becomes virtually indistinguishable from background. In evaluating the hybridization data, a threshold intensity value may be

selected below which a signal is not counted as being essentially indistinguishable from background.

Where it is desirable to detect nucleic acids expressed at lower levels, a lower threshold is chosen. Conversely, where only high expression levels are to be evaluated a higher threshold level is selected. In a preferred embodiment, a suitable threshold is about 10% above that of the average background signal.

In addition, the provision of appropriate controls permits a more detailed analysis that controls for variations in hybridization conditions, cell health, non-specific binding and the like. Thus, for example, in a preferred embodiment, the hybridization array is provided with normalization controls as described above in Section IV.A.2. These normalization controls are probes complementary to control sequences added in a known concentration to the sample. Where the overall hybridization conditions are poor, the normalization controls will show a smaller signal reflecting reduced hybridization. Conversely, where hybridization conditions are good, the normalization controls will provide a higher signal reflecting the improved hybridization. Normalization of the signal derived from other probes in the array to the normalization controls thus provides a control for variations in hybridization conditions. Typically, normalization is accomplished by dividing the measured signal from the other probes in the array by the average signal produced by the normalization controls. Normalization may also include correction for variations due to sample preparation and amplification. Such normalization may be accomplished by dividing the measured signal by the average signal from the sample preparation/amplfication control probes (*e.g.*, the Bio B probes). The resulting values may be multiplied by a constant value to scale the results.

As indicated above, the high density array can include mismatch controls. In a preferred embodiment, there is a mismatch control having a central mismatch for every probe (except the normalization controls) in the array. It is expected that after washing in stringent conditions, where a perfect match would be expected to hybridize to the probe, but not to the mismatch, the signal from the mismatch controls should only reflect non-specific binding or the presence in the sample of a nucleic acid that hybridizes with the mismatch. Where both the probe in question and its corresponding mismatch control both show high signals, or the mismatch shows a higher signal than its

corresponding test probe, there is a problem with the hybridization and the signal from those probes is ignored. The difference in hybridization signal intensity between the target specific probe and its corresponding mismatch control is a measure of the discrimination of the target-specific probe. Thus, in a preferred embodiment, the signal of the mismatch probe is subtracted from the signal from its corresponding test probe to provide a measure of the signal due to specific binding of the test probe.

The concentration of a particular sequence can then be determined by measuring the signal intensity of each of the probes that bind specifically to that gene and normalizing to the normalization controls. Where the signal from the probes is greater than the mismatch, the mismatch is subtracted. Where the mismatch intensity is equal to or greater than its corresponding test probe, the signal is ignored. The expression level of a particular gene can then be scored by the number of positive signals (either absolute or above a threshold value), the intensity of the positive signals (either absolute or above a selected threshold value), or a combination of both metrics (*e.g.*, a weighted average).

It is a surprising discovery of this invention, that normalization controls are often unnecessary for useful quantification of a hybridization signal. Thus, where optimal probes have been identified in the two step selection process as described above, in Section II.B., the average hybridization signal produced by the selected optimal probes provides a good quantified measure of the concentration of hybridized nucleic acid.

IX. Computer-implemented Expression Monitoring

The methods of monitoring gene expression of this invention may be performed utilizing a computer. The computer typically runs a software program that includes computer code incorporating the invention for analyzing hybridization intensities measured from a substrate or chip and thus, monitoring the expression of one or more genes. Although the following will describe specific embodiments of the invention, the invention is not limited to any one embodiment so the following is for purposes of illustration and not limitation.

Fig. 6 illustrates an example of a computer system used to execute the software of an embodiment of the present invention. As shown, shows a computer system 100 includes a monitor 102, screen 104, cabinet 106, keyboard 108, and mouse 110. Mouse 110 may have one or more buttons such as mouse buttons 112. Cabinet 106 houses a CD-ROM drive 114, a system memory and a hard drive (both shown in Fig. 7) which may be utilized to store and retrieve software programs incorporating computer code that implements the invention, data for use with the invention, and the like. Although a CD-ROM 116 is shown as an exemplary computer readable storage medium, other computer readable storage media including floppy disks, tape, flash memory, system memory, and hard drives may be utilized. Cabinet 106 also houses familiar computer components (not shown) such as a central processor, system memory, hard disk, and the like.

Fig. 7 shows a system block diagram of computer system 100 used to execute the software of an embodiment of the present invention. As in Fig. 6, computer system 100 includes monitor 102 and keyboard 108. Computer system 100 further includes subsystems such as a central processor 120, system memory 122, I/O controller 124, display adapter 126, removable disk 128 (*e.g.*, CD-ROM drive), fixed disk 130 (*e.g.*, hard drive), network interface 132, and speaker 134. Other computer systems suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system could include more than one processor 120 (*i.e.*, a multi-processor system) or a cache memory.

Arrows such as 136 represent the system bus architecture of computer system 100. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, a local bus could be utilized to connect the central processor to the system memory and display adapter. Computer system 100 shown in Fig. 7 is but an example of a computer system suitable for use with the present invention. Other configurations of subsystems suitable for use with the present invention will be readily apparent to one of ordinary skill in the art.

Fig. 8 shows a flowchart of a process of monitoring the expression of a gene. The process compares hybridization intensities of pairs of perfect match and mismatch probes that are preferably covalently attached to the surface of a substrate or

chip. Most preferably, the nucleic acid probes have a density greater than about 60 different nucleic acid probes per 1 cm^2 of the substrate. Although the flowcharts show a sequence of steps for clarity, this is not an indication that the steps must be performed in this specific order. One of ordinary skill in the art would readily recognize that many of the steps may be reordered, combined, and deleted without departing from the invention.

Initially, nucleic acid probes are selected that are complementary to the target sequence (or gene). These probes are the perfect match probes. Another set of probes is specified that are intended to be not perfectly complementary to the target sequence. These probes are the mismatch probes and each mismatch probe includes at least one nucleotide mismatch from a perfect match probe. Accordingly, a mismatch probe and the perfect match probe from which it was derived make up a pair of probes. As mentioned earlier, the nucleotide mismatch is preferably near the center of the mismatch probe.

The probe lengths of the perfect match probes are typically chosen to exhibit high hybridization affinity with the target sequence. For example, the nucleic acid probes may be all 20-mers. However, probes of varying lengths may also be synthesized on the substrate for any number of reasons including resolving ambiguities.

The target sequence is typically fragmented, labeled and exposed to a substrate including the nucleic acid probes as described earlier. The hybridization intensities of the nucleic acid probes is then measured and input into a computer system. The computer system may be the same system that directs the substrate hybridization or it may be a different system altogether. Of course, any computer system for use with the invention should have available other details of the experiment including possibly the gene name, gene sequence, probe sequences, probe locations on the substrate, and the like.

Referring to Fig. 8, after hybridization, the computer system receives input of hybridization intensities of the multiple pairs of perfect match and mismatch probes at step 202. The hybridization intensities indicate hybridization affinity between the nucleic acid probes and the target nucleic acid (which corresponds to a gene). Each pair includes a perfect match probe that is perfectly complementary to a portion of the

target nucleic acid and a mismatch probe that differs from the perfect match probe by at least one nucleotide.

At step 204, the computer system compares the hybridization intensities of the perfect match and mismatch probes of each pair. If the gene is expressed, the hybridization intensity (or affinity) of a perfect match probe of a pair should be recognizably higher than the corresponding mismatch probe. Generally, if the hybridizations intensities of a pair of probes are substantially the same, it may indicate the gene is not expressed. However, the determination is not based on a single pair of probes, the determination of whether a gene is expressed is based on an analysis of many pairs of probes. An exemplary process of comparing the hybridization intensities of the pairs of probes will be described in more detail in reference to Fig. 9.

After the system compares the hybridization intensity of the perfect match and mismatch probes, the system indicates expression of the gene at step 206. As an example, the system may indicate to a user that the gene is either present (expressed), marginal or absent (unexpressed).

Fig. 9 shows a flowchart of a process of determining if a gene is expressed utilizing a decision matrix. At step 252, the computer system receives raw scan data of N pairs of perfect match and mismatch probes. In a preferred embodiment, the hybridization intensities are photon counts from a fluorescein labeled target that has hybridized to the probes on the substrate. For simplicity, the hybridization intensity of a perfect match probe will be designed "I_{pm}" and the hybridization intensity of a mismatch probe will be designed "I_{mm}."

Hybridization intensities for a pair of probes is retrieved at step 254. The background signal intensity is subtracted from each of the hybridization intensities of the pair at step 256. Background subtraction may also be performed on all the raw scan data at the same time.

At step 258, the hybridization intensities of the pair of probes are compared to a difference threshold (D) and a ratio threshold (R). It is determined if the difference between the hybridization intensities of the pair ($I_{pm} - I_{mm}$) is greater than or equal to the difference threshold AND the quotient of the hybridization intensities of the pair (I_{pm} / I_{mm}) is greater than or equal to the ratio threshold. The difference thresholds

are typically user defined values that have been determined to produce accurate expression monitoring of a gene or genes. In one embodiment, the difference threshold is 20 and the ratio threshold is 1.2.

If $I_{pm} - I_{mm} >= D$ and $I_{pm} / I_{mm} >= R$, the value NPOS is incremented at step 260. In general, NPOS is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely expressed. NPOS is utilized in a determination of the expression of the gene.

At step 262, it is determined if $I_{mm} - I_{pm} >= D$ and $I_{mm} / I_{pm} >= R$. If this expression is true, the value NNEG is incremented at step 264. In general, NNEG is a value that indicates the number of pairs of probes which have hybridization intensities indicating that the gene is likely not expressed. NNEG, like NPOS, is utilized in a determination of the expression of the gene.

For each pair that exhibits hybridization intensities either indicating the gene is expressed or not expressed, a log ratio value (LR) and intensity difference value (IDIF) are calculated at step 266. LR is calculated by the log of the quotient of the hybridization intensities of the pair (I_{pm} / I_{mm}). The IDIF is calculated by the difference between the hybridization intensities of the pair ($I_{pm} - I_{mm}$). If there is a next pair of hybridization intensities at step 268, they are retrieved at step 254.

At step 272, a decision matrix is utilized to indicate if the gene is expressed. The decision matrix utilizes the values N, NPOS, NNEG, and LR (multiple LRs). The following four assignments are performed:

P1 = NPOS / NNEG

P2 = NPOS / N

P3 = (10 * SUM(LR)) / (NPOS + NNEG)

These P values are then utilized to determine if the gene is expressed.

For purposes of illustration, the P values are broken down into ranges. If P1 is greater than or equal to 2.1, then A is true. If P1 is less than 2.1 and greater than or equal to 1.8, then B is true. Otherwise, C is true. Thus, P1 is broken down into three ranges A, B and C. This is done to aid the readers understanding of the invention.

Thus, all of the P values are broken down into ranges according to the following:

A = (P1 >= 2.1)

B = (2.1 > P1 >= 1.8)

C = (P1 < 1.8)

X = (P2 >= 0.35)

Y = (0.35 > P2 >= 0.20)

Z = (P2 < 0.20)

Q = (P3 >= 1.5)

R = (1.5 > P3 >= 1.1)

S = (P3 < 1.1)

Once the P values are broken down into ranges according to the above boolean values, the gene expression is determined.

The gene expression is indicated as present (expressed), marginal or absent (not expressed). The gene is indicated as expressed if the following expression is true: A and (X or Y) and (Q or R). In other words, the gene is indicated as expressed if P1 >= 2.1, P2 >= 0.20 and P3 >= 1.1. Additionally, the gene is indicated as expressed if the following expression is true: B and X and Q.

With the forgoing explanation, the following is a summary of the gene expression indications:

Present	A and (X or Y) and (Q or R)
	B and X and I
Marginal	A and X and S
	B and X and R
	B and Y and (Q or R)
Absent	All others cases (*e.g.*, any C combination)

In the output to the user, present may be indicated as "P," marginal as "M" and absent as "A" at step 274.

Once all the pairs of probes have been processed and the expression of the gene indicated, an average of ten times the LRs is computed at step 275. Additionally, an average of the IDIF values for the probes that incremented NPOS and NNEG is calculated. These values may be utilized for quantitative comparisons of this experiments with other experiments.

Quantitative measurements may be performed at step 276. For example, the current experiment may be compared to a previous experiment (*e.g.*, utilizing values calculated at step 270). Additionally, the experiment may be compared to hybridization intensities of RNA (such as from bacteria) present in the biological sample in a known quantity. In this manner, one may verify the correctness of the gene expression indication or call, modify threshold values, or perform any number of modifications of the preceding.

For simplicity, Fig. 9 was described in reference to a single gene. However, the process may be utilized on multiple genes in a biological sample. Therefore, any discussion of the analysis of a single gene is not an indication that the process may not be extended to processing multiple genes.

Figs. 10A and 10B show the flow of a process of determining the expression of a gene by comparing baseline scan data and experimental scan data. For example, the baseline scan data may be from a biological sample where it is known the gene is expressed. Thus, this scan data may be compared to a different biological sample to determine if the gene is expressed. Additionally, it may be determined how the expression of a gene or genes changes over time in a biological organism.

At step 302, the computer system receives raw scan data of N pairs of perfect match and mismatch probes from the baseline. The hybridization intensity of a perfect match probe from the baseline will be designed "I_{pm}" and the hybridization intensity of a mismatch probe from the baseline will be designed "I_{mm}." The background signal intensity is subtracted from each of the hybridization intensities of the pairs of baseline scan data at step 304.

At step 306, the computer system receives raw scan data of N pairs of perfect match and mismatch probes from the experimental biological sample. The hybridization intensity of a perfect match probes from the experiment will be designed

"J_{pm}" and the hybridization intensity of a mismatch probe from the experiment will be designed "J_{mm}." The background signal intensity is subtracted from each of the hybridization intensities of the pairs of experimental scan data at step 308.

The hybridization intensities of an I and J pair may be normalized at step 310. For example, the hybridization intensities of the I and J pairs may be divided by the hybridization intensity of control probes as discussed in Section II.A.2.

At step 312, the hybridization intensities of the I and J pair of probes are compared to a difference threshold (DDIF) and a ratio threshold (RDIF). It is determined if the difference between the hybridization intensities of the one pair (J_{pm} - J_{mm}) and the other pair (I_{pm} - I_{mm}) are greater than or equal to the difference threshold AND the quotient of the hybridization intensities of one pair (J_{pm} - J_{mm}) and the other pair (I_{pm} - I_{mm}) are greater than or equal to the ratio threshold. The difference thresholds are typically user defined values that have been determined to produce accurate expression monitoring of a gene or genes.

If (J_{pm} - J_{mm}) - (I_{pm} - I_{mm}) >= DDIF and (J_{pm} - J_{mm}) / (I_{pm} - I_{mm}) >= RDIF, the value NINC is incremented at step 314. In general, NINC is a value that indicates the experimental pair of probes indicates that the gene expression is likely greater (or increased) than the baseline sample. NINC is utilized in a determination of whether the expression of the gene is greater (or increased), less (or decreased) or did not change in the experimental sample compared to the baseline sample.

At step 316, it is determined if (J_{pm} - J_{mm}) - (I_{pm} - I_{mm}) >= DDIF and (J_{pm} - J_{mm}) / (I_{pm} / I_{mm}) >= RDIF. If this expression is true, NDEC is incremented. In general, NDEC is a value that indicates the experimental pair of probes indicates that the gene expression is likely less (or decreased) than the baseline sample. NDEC is utilized in a determination of whether the expression of the gene is greater (or increased), less (or decreased) or did not change in the experimental sample compared to the baseline sample.

For each of the pairs that exhibits hybridization intensities either indicating the gene is expressed more or less in the experimental sample, the values NPOS, NNEG and LR are calculated for each pair of probes. These values are calculated as discussed above in reference to Fig. 9. A suffix of either "B" or "E" has

been added to each value in order to indicate if the value denotes the baseline sample or the experimental sample, respectively. If there are next pairs of hybridization intensities at step 322, they are processed in a similar manner as shown.

Referring now to Fig. 10B, an absolute decision computation is performed for both the baseline and experimental samples at step 324. The absolute decision computation is an indication of whether the gene is expressed, marginal or absent in each of the baseline and experimental samples. Accordingly, in a preferred embodiment, this step entails performing steps 272 and 274 from Fig. 9 for each of the samples. This being done, there is an indication of gene expression for each of the samples taken alone.

At step 326, a decision matrix is utilized to determine the difference in gene expression between the two samples. This decision matrix utilizes the values, N, NPOSB, NPOSE, NNEGB, NNEGE, NINC, NDEC, LRB, and LRE as they were calculated above. The decision matrix performs different calculations depending on whether NINC is greater than or equal to NDEC. The calculations are as follows.

If NINC > = NDEC, the following four P values are determined:

P1 = NINC / NDEC

P2 = NINC / N

P3 = ((NPOSE - NPOSB) - (NNEGE - NNEGB)) / N

P4 = 10 * SUM(LRE - LRB) / N

These P values are then utilized to determine the difference in gene expression between the two samples.

For purposes of illustration, the P values are broken down into ranges as was done previously. Thus, all of the P values are broken down into ranges according to the following: